Tenth Edition

ERWIN KREYSZIG

ADVANCED ENGINEERING MATHEMATICS

Systems of Units. Some Important Conversion Factors

The most important systems of units are shown in the table below. The mks system is also known as the *International System of Units* (abbreviated *SI*), and the abbreviations sec (instead of s), gm (instead of g), and nt (instead of N) are also used.

System of units	Length	Mass	Time	Force
cgs system	centimeter (cm)	gram (g)	second (s)	dyne
mks system	meter (m)	kilogram (kg)	second (s)	newton (nt)
Engineering system	foot (ft)	slug	second (s)	pound (lb)

1 inch (in.) = 2.540000 cm

1 foot (ft) = 12 in. = 30.480000 cm

1 yard (yd) = 3 ft = 91.440000 cm

1 statute mile (mi) = 5280 ft = 1.609344 km

1 nautical mile = 6080 ft = 1.853184 km

1 acre = 4840 yd^2 = 4046.8564 m^2

1 mi^2 = 640 acres = 2.5899881 km^2

1 fluid ounce = 1/128 U.S. gallon = 231/128 $in.^3$ = 29.573730 cm^3

1 U.S. gallon = 4 quarts (liq) = 8 pints (liq) = 128 fl oz = 3785.4118 cm^3

1 British Imperial and Canadian gallon = 1.200949 U.S. gallons = 4546.087 cm^3

1 slug = 14.59390 kg

1 pound (lb) = 4.448444 nt

1 newton (nt) = 10^5 dynes

1 British thermal unit (Btu) = 1054.35 joules

1 joule = 10^7 ergs

1 calorie (cal) = 4.1840 joules

1 kilowatt-hour (kWh) = 3414.4 Btu = $3.6 \cdot 10^6$ joules

1 horsepower (hp) = 2542.48 Btu/h = 178.298 cal/sec = 0.74570 kW

1 kilowatt (kW) = 1000 watts = 3414.43 Btu/h = 238.662 cal/s

$°F = °C \cdot 1.8 + 32$

$1° = 60' = 3600'' = 0.017453293$ radian

For further details see, for example, D. Halliday, R. Resnick, and J. Walker, *Fundamentals of Physics.* 9th ed., Hoboken, N. J: Wiley, 2011. See also AN American National Standard, ASTM/IEEE Standard Metric Practice, Institute of Electrical and Electronics Engineers, Inc. (IEEE), 445 Hoes Lane, Piscataway, N. J. 08854, website at www.ieee.org.

Differentiation

$(cu)' = cu' \qquad (c \text{ constant})$

$(u + v)' = u' + v'$

$(uv)' = u'v + uv'$

$\left(\dfrac{u}{v}\right)' = \dfrac{u'v - uv'}{v^2}$

$\dfrac{du}{dx} = \dfrac{du}{dy} \cdot \dfrac{dy}{dx} \qquad \text{(Chain rule)}$

$(x^n)' = nx^{n-1}$

$(e^x)' = e^x$

$(e^{ax})' = ae^{ax}$

$(a^x)' = a^x \ln a$

$(\sin x)' = \cos x$

$(\cos x)' = -\sin x$

$(\tan x)' = \sec^2 x$

$(\cot x)' = -\csc^2 x$

$(\sinh x)' = \cosh x$

$(\cosh x)' = \sinh x$

$(\ln x)' = \dfrac{1}{x}$

$(\log_a x)' = \dfrac{\log_a e}{x}$

$(\arcsin x)' = \dfrac{1}{\sqrt{1 - x^2}}$

$(\arccos x)' = -\dfrac{1}{\sqrt{1 - x^2}}$

$(\arctan x)' = \dfrac{1}{1 + x^2}$

$(\text{arccot}\, x)' = -\dfrac{1}{1 + x^2}$

Integration

$\int uv'\, dx = uv - \int u'v\, dx$ (by parts)

$\int x^n\, dx = \dfrac{x^{n+1}}{n + 1} + c \qquad (n \neq -1)$

$\int \dfrac{1}{x}\, dx = \ln |x| + c$

$\int e^{ax}\, dx = \dfrac{1}{a} e^{ax} + c$

$\int \sin x\, dx = -\cos x + c$

$\int \cos x\, dx = \sin x + c$

$\int \tan x\, dx = -\ln |\cos x| + c$

$\int \cot x\, dx = \ln |\sin x| + c$

$\int \sec x\, dx = \ln |\sec x + \tan x| + c$

$\int \csc x\, dx = \ln |\csc x - \cot x| + c$

$\int \dfrac{dx}{x^2 + a^2} = \dfrac{1}{a} \arctan \dfrac{x}{a} + c$

$\int \dfrac{dx}{\sqrt{a^2 - x^2}} = \arcsin \dfrac{x}{a} + c$

$\int \dfrac{dx}{\sqrt{x^2 + a^2}} = \text{arcsinh}\, \dfrac{x}{a} + c$

$\int \dfrac{dx}{\sqrt{x^2 - a^2}} = \text{arccosh}\, \dfrac{x}{a} + c$

$\int \sin^2 x\, dx = \frac{1}{2}x - \frac{1}{4} \sin 2x + c$

$\int \cos^2 x\, dx = \frac{1}{2}x + \frac{1}{4} \sin 2x + c$

$\int \tan^2 x\, dx = \tan x - x + c$

$\int \cot^2 x\, dx = -\cot x - x + c$

$\int \ln x\, dx = x \ln x - x + c$

$$\int e^{ax} \sin bx\, dx = \frac{e^{ax}}{a^2 + b^2}(a \sin bx - b \cos bx) + c$$

$$\int e^{ax} \cos bx\, dx = \frac{e^{ax}}{a^2 + b^2}(a \cos bx + b \sin bx) + c$$

ADVANCED ENGINEERING MATHEMATICS

10TH EDITION

ADVANCED ENGINEERING MATHEMATICS

ERWIN KREYSZIG

Professor of Mathematics
Ohio State University
Columbus, Ohio

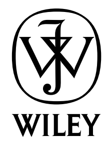
In collaboration with

HERBERT KREYSZIG

New York, New York

EDWARD J. NORMINTON

Associate Professor of Mathematics
Carleton University
Ottawa, Ontario



JOHN WILEY & SONS, INC.

PUBLISHER	Laurie Rosatone
PROJECT EDITOR	Shannon Corliss
MARKETING MANAGER	Jonathan Cottrell
CONTENT MANAGER	Lucille Buonocore
PRODUCTION EDITOR	Barbara Russiello
MEDIA EDITOR	Melissa Edwards
MEDIA PRODUCTION SPECIALIST	Lisa Sabatini
TEXT AND COVER DESIGN	Madelyn Lesure
PHOTO RESEARCHER	Sheena Goldstein
COVER PHOTO	© Denis Jr. Tangney/iStockphoto Cover photo shows the Zakim Bunker Hill Memorial Bridge in Boston, MA.

This book was set in Times Roman. The book was composed by PreMedia Global, and printed and bound by RR Donnelley & Sons Company, Jefferson City, MO. The cover was printed by RR Donnelley & Sons Company, Jefferson City, MO.

This book is printed on acid free paper. ∞

Founded in 1807, John Wiley & Sons, Inc. has been a valued source of knowledge and understanding for more than 200 years, helping people around the world meet their needs and fulfill their aspirations. Our company is built on a foundation of principles that include responsibility to the communities we serve and where we live and work. In 2008, we launched a Corporate Citizenship Initiative, a global effort to address the environmental, social, economic, and ethical challenges we face in our business. Among the issues we are addressing are carbon impact, paper specifications and procurement, ethical conduct within our business and among our vendors, and community and charitable support. For more information, please visit our website: www.wiley.com/go/citizenship.

ISBN 978-0-470-45836-5

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

PREFACE

See also http://www.wiley.com/college/kreyszig

Purpose and Structure of the Book

This book provides a comprehensive, thorough, and up-to-date treatment of ***engineering mathematics***. It is intended to introduce students of engineering, physics, mathematics, computer science, and related fields to those areas of ***applied mathematics*** that are most relevant for solving practical problems. A course in elementary calculus is the sole ***prerequisite***. (However, a concise refresher of basic calculus for the student is included on the inside cover and in Appendix 3.)

The subject matter is arranged into seven parts as follows:

A. Ordinary Differential Equations (ODEs) in Chapters 1–6
B. Linear Algebra. Vector Calculus. See Chapters 7–10
C. Fourier Analysis. Partial Differential Equations (PDEs). See Chapters 11 and 12
D. Complex Analysis in Chapters 13–18
E. Numeric Analysis in Chapters 19–21
F. Optimization, Graphs in Chapters 22 and 23
G. Probability, Statistics in Chapters 24 and 25.

These are followed by five appendices: **1.** References, **2.** Answers to Odd-Numbered Problems, **3.** Auxiliary Materials (see also inside covers of book), **4.** Additional Proofs, **5.** Table of Functions. This is shown in a block diagram on the next page.

The parts of the book are kept independent. In addition, individual chapters are kept as independent as possible. (If so needed, any prerequisites—to the level of individual sections of prior chapters—are clearly stated at the opening of each chapter.) We give the instructor **maximum flexibility in selecting the material** and tailoring it to his or her need. *The book has helped to pave the way for the present development of engineering mathematics*. This new edition will prepare the student for the current tasks and the future by a modern approach to the areas listed above. We provide the material and learning tools for the students to get a good foundation of engineering mathematics that will help them in their careers and in further studies.

General Features of the Book Include:

- **Simplicity of examples** to make the book teachable—why choose complicated examples when simple ones are as instructive or even better?
- **Independence of parts and blocks of chapters** to provide flexibility in tailoring courses to specific needs.
- **Self-contained presentation**, except for a few clearly marked places where a proof would exceed the level of the book and a reference is given instead.
- **Gradual increase in difficulty of material with no jumps or gaps** to ensure an enjoyable teaching and learning experience.
- **Modern standard notation** to help students with other courses, modern books, and journals in mathematics, engineering, statistics, physics, computer science, and others.

Furthermore, we designed the book to be a **single, self-contained, authoritative, and convenient source** for studying and teaching applied mathematics, eliminating the need for time-consuming searches on the Internet or time-consuming trips to the library to get a particular reference book.

PARTS AND CHAPTERS OF THE BOOK

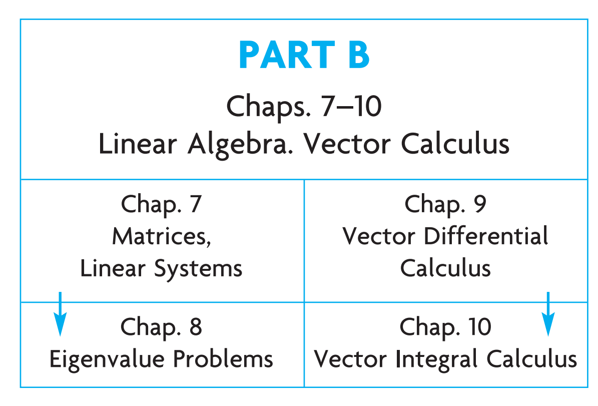
PART A

Chaps. 1–6
Ordinary Differential Equations (ODEs)

Chaps. 1–4
Basic Material

Chap. 5
Series Solutions

Chap. 6
Laplace Transforms



PART B

Chaps. 7–10
Linear Algebra. Vector Calculus

Chap. 7
Matrices, Linear Systems

Chap. 9
Vector Differential Calculus

Chap. 8
Eigenvalue Problems

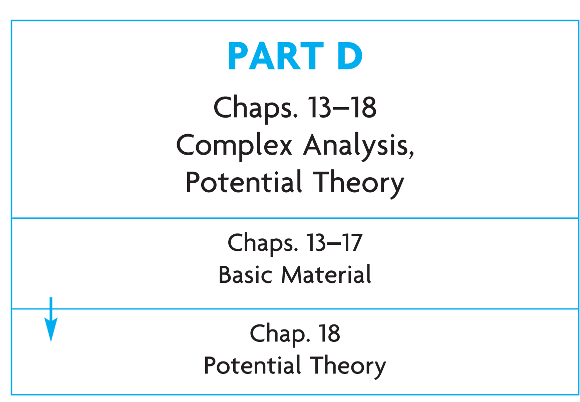
Chap. 10
Vector Integral Calculus

PART C

Chaps. 11–12
Fourier Analysis. Partial Differential Equations (PDEs)

Chap. 11
Fourier Analysis

Chap. 12
Partial Differential Equations



PART D

Chaps. 13–18
Complex Analysis, Potential Theory

Chaps. 13–17
Basic Material

Chap. 18
Potential Theory

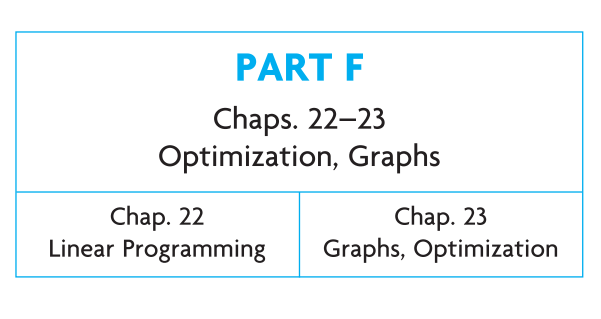
PART E

Chaps. 19–21
Numeric Analysis

Chap. 19
Numerics in General

Chap. 20
Numeric Linear Algebra

Chap. 21
Numerics for ODEs and PDEs



PART F

Chaps. 22–23
Optimization, Graphs

Chap. 22
Linear Programming

Chap. 23
Graphs, Optimization

PART G

Chaps. 24–25
Probability, Statistics

Chap. 24
Data Analysis. Probability Theory

Chap. 25
Mathematical Statistics

GUIDES AND MANUALS

Maple Computer Guide
Mathematica Computer Guide

Student Solutions Manual
and Study Guide

Instructor's Manual

Four Underlying Themes of the Book

The driving force in engineering mathematics is the rapid growth of technology and the sciences. New areas—often drawing from several disciplines—come into existence. Electric cars, solar energy, wind energy, green manufacturing, nanotechnology, risk management, biotechnology, biomedical engineering, computer vision, robotics, space travel, communication systems, green logistics, transportation systems, financial engineering, economics, and many other areas are advancing rapidly. What does this mean for engineering mathematics? The engineer has to take a problem from any diverse area and be able to model it. This leads to the first of four underlying themes of the book.

1. Modeling is the process in engineering, physics, computer science, biology, chemistry, environmental science, economics, and other fields whereby a physical situation or some other observation is translated into a mathematical model. This mathematical model could be a system of differential equations, such as in population control (Sec. 4.5), a probabilistic model (Chap. 24), such as in risk management, a linear programming problem (Secs. 22.2–22.4) in minimizing environmental damage due to pollutants, a financial problem of valuing a bond leading to an algebraic equation that has to be solved by Newton's method (Sec. 19.2), and many others.

The next step is **solving the mathematical problem** obtained by one of the many techniques covered in *Advanced Engineering Mathematics*.

The third step is **interpreting the mathematical result** in physical or other terms to see what it means in practice and any implications.

Finally, we may have to **make a decision** that may be of an industrial nature or **recommend a public policy**. For example, the population control model may imply the policy to stop fishing for 3 years. Or the valuation of the bond may lead to a recommendation to buy. The variety is endless, but the underlying mathematics is surprisingly powerful and able to provide advice leading to the achievement of goals toward the betterment of society, for example, by recommending wise policies concerning global warming, better allocation of resources in a manufacturing process, or making statistical decisions (such as in Sec. 25.4 whether a drug is effective in treating a disease).

While we cannot predict what the future holds, we do know that the student has to practice modeling by being given problems from many different applications as is done in this book. We teach modeling from scratch, right in Sec. 1.1, and give many examples in Sec. 1.3, and continue to reinforce the modeling process throughout the book.

2. Judicious use of powerful software for numerics (listed in the beginning of Part E) and statistics (Part G) is of growing importance. Projects in engineering and industrial companies may involve large problems of modeling very complex systems with hundreds of thousands of equations or even more. They require the use of such software. However, our policy has always been to leave it up to the instructor to determine the degree of use of computers, from none or little use to extensive use. More on this below.

3. The beauty of engineering mathematics. ***Engineering mathematics relies on relatively few basic concepts and involves powerful unifying principles***. We point them out whenever they are clearly visible, such as in Sec. 4.1 where we "grow" a mixing problem from one tank to two tanks and a circuit problem from one circuit to two circuits, thereby also increasing the number of ODEs from one ODE to two ODEs. This is an example of an attractive mathematical model because the "growth" in the problem is reflected by an "increase" in ODEs.

4. To clearly identify the conceptual structure of subject matters. For example, complex analysis (in Part D) is a field that is not monolithic in structure but was formed by three distinct schools of mathematics. Each gave a different approach, which we clearly mark. The first approach is solving complex integrals by Cauchy's integral formula (Chaps. 13 and 14), the second approach is to use the Laurent series and solve complex integrals by residue integration (Chaps. 15 and 16), and finally we use a geometric approach of conformal mapping to solve boundary value problems (Chaps. 17 and 18). Learning the conceptual structure and terminology of the different areas of engineering mathematics is very important for three reasons:

a. It allows the student to *identify a new problem and put it into the right group of problems*. The areas of engineering mathematics are growing but most often retain their conceptual structure.

b. The student can *absorb new information more rapidly* by being able to fit it into the conceptual structure.

c. Knowledge of the conceptual structure and terminology is also important when *using the Internet to search for mathematical information*. Since the search proceeds by putting in key words (i.e., terms) into the search engine, the student has to remember the important concepts (or be able to look them up in the book) that identify the application and area of engineering mathematics.

Big Changes in This Edition

1 Problem Sets Changed

The problem sets have been revised and rebalanced with some problem sets having more problems and some less, reflecting changes in engineering mathematics. There is a greater emphasis on modeling. Now there are also problems on the discrete Fourier transform (in Sec. 11.9).

2 Series Solutions of ODEs, Special Functions and Fourier Analysis Reorganized

Chap. 5, on series solutions of ODEs and special functions, has been shortened. Chap. 11 on Fourier Analysis now contains Sturm–Liouville problems, orthogonal functions, and orthogonal eigenfunction expansions (Secs. 11.5, 11.6), where they fit better conceptually (rather than in Chap. 5), being extensions of Fourier's idea of using orthogonal functions.

3 Openings of Parts and Chapters Rewritten As Well As Parts of Sections

In order to give the student a better idea of the structure of the material (see Underlying Theme 4 above), we have entirely rewritten the openings of parts and chapters. Furthermore, large parts or individual paragraphs of sections have been rewritten or new sentences inserted into the text. This should give the students a better intuitive understanding of the material (see Theme 3 above), let them draw conclusions on their own, and be able to tackle more advanced material. Overall, we feel that the book has become more detailed and leisurely written.

4 Student Solutions Manual and Study Guide Enlarged

Upon the explicit request of the users, the answers provided are more detailed and complete. More explanations are given on how to learn the material effectively by pointing out what is most important.

5 More Historical Footnotes, Some Enlarged

Historical footnotes are there to show the student that many people from different countries working in different professions, such as surveyors, researchers in industry, etc., contributed

to the field of engineering mathematics. It should encourage the students to be creative in their own interests and careers and perhaps also to make contributions to engineering mathematics.

Further Changes and New Features

- Parts of Chap. 1 on first-order ODEs are rewritten. More emphasis on modeling, also new block diagram explaining this concept in Sec. 1.1. Early introduction of Euler's method in Sec. 1.2 to familiarize student with basic numerics. More examples of separable ODEs in Sec. 1.3.
- For Chap. 2, on second-order ODEs, note the following changes: For ease of reading, the first part of Sec. 2.4, which deals with setting up the mass-spring system, has been rewritten; also some rewriting in Sec. 2.5 on the Euler–Cauchy equation.
- Substantially shortened Chap. 5, Series Solutions of ODEs. Special Functions: combined Secs. 5.1 and 5.2 into one section called "Power Series Method," shortened material in Sec. 5.4 Bessel's Equation (of the first kind), removed Sec. 5.7 (Sturm–Liouville Problems) and Sec. 5.8 (Orthogonal Eigenfunction Expansions) and moved material into Chap. 11 (see "Major Changes" above).
- New equivalent definition of ***basis*** (Sec. 7.4).
- In Sec. 7.9, completely new part on **composition of linear transformations** with two new examples. Also, more detailed explanation of the role of axioms, in connection with the definition of vector space.
- New table of orientation (opening of Chap. 8 "Linear Algebra: Matrix Eigenvalue Problems") where eigenvalue problems occur in the book. More intuitive explanation of what an eigenvalue is at the begining of Sec. 8.1.
- Better definition of ***cross product*** (in vector differential calculus) by properly identifying the degenerate case (in Sec. 9.3).
- **Chap. 11 on Fourier Analysis extensively rearranged:** Secs. 11.2 and 11.3 combined into one section (Sec. 11.2), old Sec. 11.4 on complex Fourier Series removed and new Secs. 11.5 (Sturm–Liouville Problems) and 11.6 (Orthogonal Series) put in (see "Major Changes" above). New problems (new!) in problem set 11.9 on **discrete Fourier transform**.
- **New section 12.5** on modeling heat flow from a body in space by setting up the heat equation. Modeling PDEs is more difficult so we separated the modeling process from the solving process (in Sec. 12.6).
- **Introduction to Numerics** rewritten for greater clarity and better presentation; new Example 1 on how to round a number. Sec. 19.3 on interpolation shortened by removing the less important central difference formula and giving a reference instead.
- Large new footnote with historical details in Sec. 22.3, honoring George Dantzig, the inventor of the **simplex method**.
- **Traveling salesman problem** now described better as a "difficult" problem, typical of combinatorial optimization (in Sec. 23.2). More careful explanation on how to compute the capacity of a cut set in Sec. 23.6 (Flows on Networks).
- In Chap. 24, material on data representation and characterization restructured in terms of five examples and enlarged to include empirical rule on distribution of

data, outliers, and the z-score (Sec. 24.1). Furthermore, new example on encription (Sec. 24.4).

- Lists of **software** for numerics (Part E) and statistics (Part G) updated.
- References in **Appendix 1** updated to include new editions and some references to websites.

Use of Computers

The presentation in this book is ***adaptable to various degrees of use of software, Computer Algebra Systems (CAS's), or programmable graphic calculators***, ranging from no use, very little use, medium use, to intensive use of such technology. The choice of how much computer content the course should have is left up to the instructor, thereby exhibiting our philosophy of maximum flexibility and adaptability. And, no matter what the instructor decides, there will be no gaps or jumps in the text or problem set. Some problems are clearly designed as routine and drill exercises and should be solved by hand (paper and pencil, or typing on your computer). Other problems require more thinking and can also be solved without computers. Then there are problems where the computer can give the student a hand. And finally, the book has ***CAS projects***, ***CAS problems*** and ***CAS experiments***, which ***do require*** a computer, and show its power in solving problems that are difficult or impossible to access otherwise. Here our goal is to combine intelligent computer use with high-quality mathematics. The computer invites visualization, experimentation, and independent discovery work. In summary, the high degree of flexibility of computer use for the book is possible since there are plenty of problems to choose from and the CAS problems can be omitted if desired.

Note that ***information on software*** (what is available and where to order it) is at the beginning of Part E on Numeric Analysis and Part G on Probability and Statistics. Since ***Maple*** and ***Mathematica*** are popular Computer Algebra Systems, there are two computer guides available that are specifically tailored to *Advanced Engineering Mathematics:* E. Kreyszig and E.J. Norminton, *Maple Computer Guide, 10th Edition* and *Mathematica Computer Guide, 10th Edition*. Their use is completely optional as the text in the book is written without the guides in mind.

Suggestions for Courses: A Four-Semester Sequence

The material, when taken in sequence, is suitable for four consecutive semester courses, meeting 3 to 4 hours a week:

1st Semester	ODEs (Chaps. 1–5 or 1–6)
2nd Semester	Linear Algebra. Vector Analysis (Chaps. 7–10)
3rd Semester	Complex Analysis (Chaps. 13–18)
4th Semester	Numeric Methods (Chaps. 19–21)

Suggestions for Independent One-Semester Courses

The book is also suitable for various independent one-semester courses meeting 3 hours a week. For instance,

Introduction to ODEs (Chaps. 1–2, 21.1)
Laplace Transforms (Chap. 6)
Matrices and Linear Systems (Chaps. 7–8)

Vector Algebra and Calculus (Chaps. 9–10)
Fourier Series and PDEs (Chaps. 11–12, Secs. 21.4–21.7)
Introduction to Complex Analysis (Chaps. 13–17)
Numeric Analysis (Chaps. 19, 21)
Numeric Linear Algebra (Chap. 20)
Optimization (Chaps. 22–23)
Graphs and Combinatorial Optimization (Chap. 23)
Probability and Statistics (Chaps. 24–25)

Acknowledgments

We are indebted to former teachers, colleagues, and students who helped us directly or indirectly in preparing this book, in particular this new edition. We profited greatly from discussions with engineers, physicists, mathematicians, computer scientists, and others, and from their written comments. We would like to mention in particular Professors Y. A. Antipov, R. Belinski, S. L. Campbell, R. Carr, P. L. Chambré, Isabel F. Cruz, Z. Davis, D. Dicker, L. D. Drager, D. Ellis, W. Fox, A. Goriely, R. B. Guenther, J. B. Handley, N. Harbertson, A. Hassen, V. W. Howe, H. Kuhn, K. Millet, J. D. Moore, W. D. Munroe, A. Nadim, B. S. Ng, J. N. Ong, P. J. Pritchard, W. O. Ray, L. F. Shampine, H. L. Smith, Roberto Tamassia, A. L. Villone, H. J. Weiss, A. Wilansky, Neil M. Wigley, and L. Ying; Maria E. and Jorge A. Miranda, JD, all from the United States; Professors Wayne H. Enright, Francis. L. Lemire, James J. Little, David G. Lowe, Gerry McPhail, Theodore S. Norvell, and R. Vaillancourt; Jeff Seiler and David Stanley, all from Canada; and Professor Eugen Eichhorn, Gisela Heckler, Dr. Gunnar Schroeder, and Wiltrud Stiefenhofer from Europe. Furthermore, we would like to thank Professors John B. Donaldson, Bruce C. N. Greenwald, Jonathan L. Gross, Morris B. Holbrook, John R. Kender, and Bernd Schmitt; and Nicholaiv Villalobos, all from Columbia University, New York; as well as Dr. Pearl Chang, Chris Gee, Mike Hale, Joshua Jayasingh, MD, David Kahr, Mike Lee, R. Richard Royce, Elaine Schattner, MD, Raheel Siddiqui, Robert Sullivan, MD, Nancy Veit, and Ana M. Kreyszig, JD, all from New York City. We would also like to gratefully acknowledge the use of facilities at Carleton University, Ottawa, and Columbia University, New York.

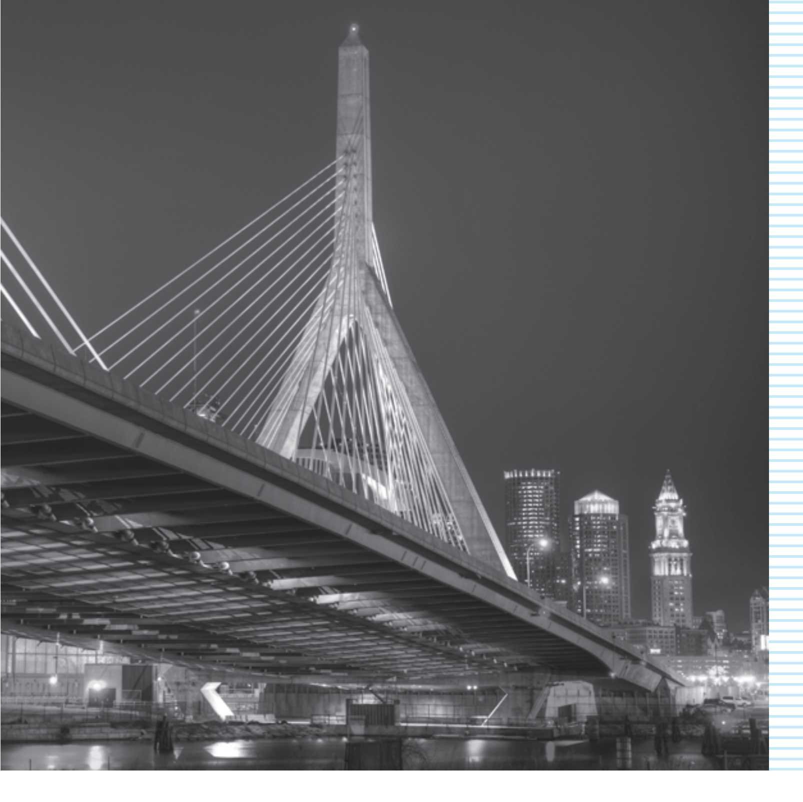
Furthermore we wish to thank John Wiley and Sons, in particular Publisher Laurie Rosatone, Editor Shannon Corliss, Production Editor Barbara Russiello, Media Editor Melissa Edwards, Text and Cover Designer Madelyn Lesure, and Photo Editor Sheena Goldstein for their great care and dedication in preparing this edition. In the same vein, we would also like to thank Beatrice Ruberto, copy editor and proofreader, WordCo, for the Index, and Joyce Franzen of PreMedia and those of PreMedia Global who typeset this edition.

Suggestions of many readers worldwide were evaluated in preparing this edition. Further comments and suggestions for improving the book will be gratefully received.

KREYSZIG

CONTENTS

PART A Ordinary Differential Equations (ODEs) 1



PART A

Ordinary Differential Equations (ODEs)

Many physical laws and relations can be expressed mathematically in the form of differential equations. Thus it is natural that this book opens with the study of differential equations and their solutions. Indeed, many engineering problems appear as differential equations.

The main objectives of Part A are twofold: the study of ordinary differential equations and their most important methods for solving them and the study of modeling.

Ordinary differential equations **(ODEs)** are differential equations that depend on a single variable. The more difficult study of partial differential equations (PDEs), that is, differential equations that depend on several variables, is covered in Part C.

Modeling is a crucial general process in engineering, physics, computer science, biology, medicine, environmental science, chemistry, economics, and other fields that translates a physical situation or some other observations into a "mathematical model." Numerous examples from engineering (e.g., mixing problem), physics (e.g., Newton's law of cooling), biology (e.g., Gompertz model), chemistry (e.g., radiocarbon dating), environmental science (e.g., population control), etc. shall be given, whereby this process is explained in detail, that is, how to set up the problems correctly in terms of differential equations.

For those interested in solving ODEs numerically on the computer, look at Secs. 21.1–21.3 of Chapter 21 of Part F, that is, ***numeric methods for ODEs***. These sections are kept independent by design of the other sections on numerics. ***This allows for the study of numerics for ODEs directly after Chap. 1 or 2.***

CHAPTER 1

First-Order ODEs

Chapter 1 begins the study of ordinary differential equations (ODEs) by deriving them from physical or other problems (***modeling***), solving them by standard mathematical methods, and interpreting solutions and their graphs in terms of a given problem. The simplest ODEs to be discussed are ODEs ***of the first order*** because they involve only the first derivative of the unknown function and no higher derivatives. These unknown functions will usually be denoted by $y(x)$ or $y(t)$ when the independent variable denotes time t. The chapter ends with a study of the existence and uniqueness of solutions of ODEs in Sec. 1.7.

Understanding the basics of ODEs requires solving problems by hand (paper and pencil, or typing on your computer, but first without the aid of a CAS). In doing so, you will gain an important conceptual understanding and feel for the basic terms, such as ODEs, direction field, and initial value problem. If you wish, you can use your **Computer Algebra System (CAS)** for checking solutions.

COMMENT. ***Numerics for first-order ODEs can be studied immediately after this chapter.*** See Secs. 21.1–21.2, which are independent of other sections on numerics.

Prerequisite: Integral calculus.
Sections that may be omitted in a shorter course: 1.6, 1.7.
References and Answers to Problems: App. 1 Part A, and App. 2.

1.1 Basic Concepts. Modeling

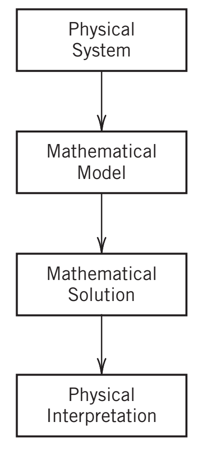


Fig. 1. Modeling, solving, interpreting

If we want to solve an engineering problem (usually of a physical nature), we first have to formulate the problem as a mathematical expression in terms of variables, functions, and equations. Such an expression is known as a mathematical **model** of the given problem. The process of setting up a model, solving it mathematically, and interpreting the result in physical or other terms is called *mathematical modeling* or, briefly, **modeling**.

Modeling needs experience, which we shall gain by discussing various examples and problems. (Your computer may often help you in *solving* but rarely in *setting up* models.)

Now many physical concepts, such as velocity and acceleration, are derivatives. Hence a model is very often an equation containing derivatives of an unknown function. Such a model is called a **differential equation**. Of course, we then want to find a solution (a function that satisfies the equation), explore its properties, graph it, find values of it, and interpret it in physical terms so that we can understand the behavior of the physical system in our given problem. However, before we can turn to methods of solution, we must first define some basic concepts needed throughout this chapter.

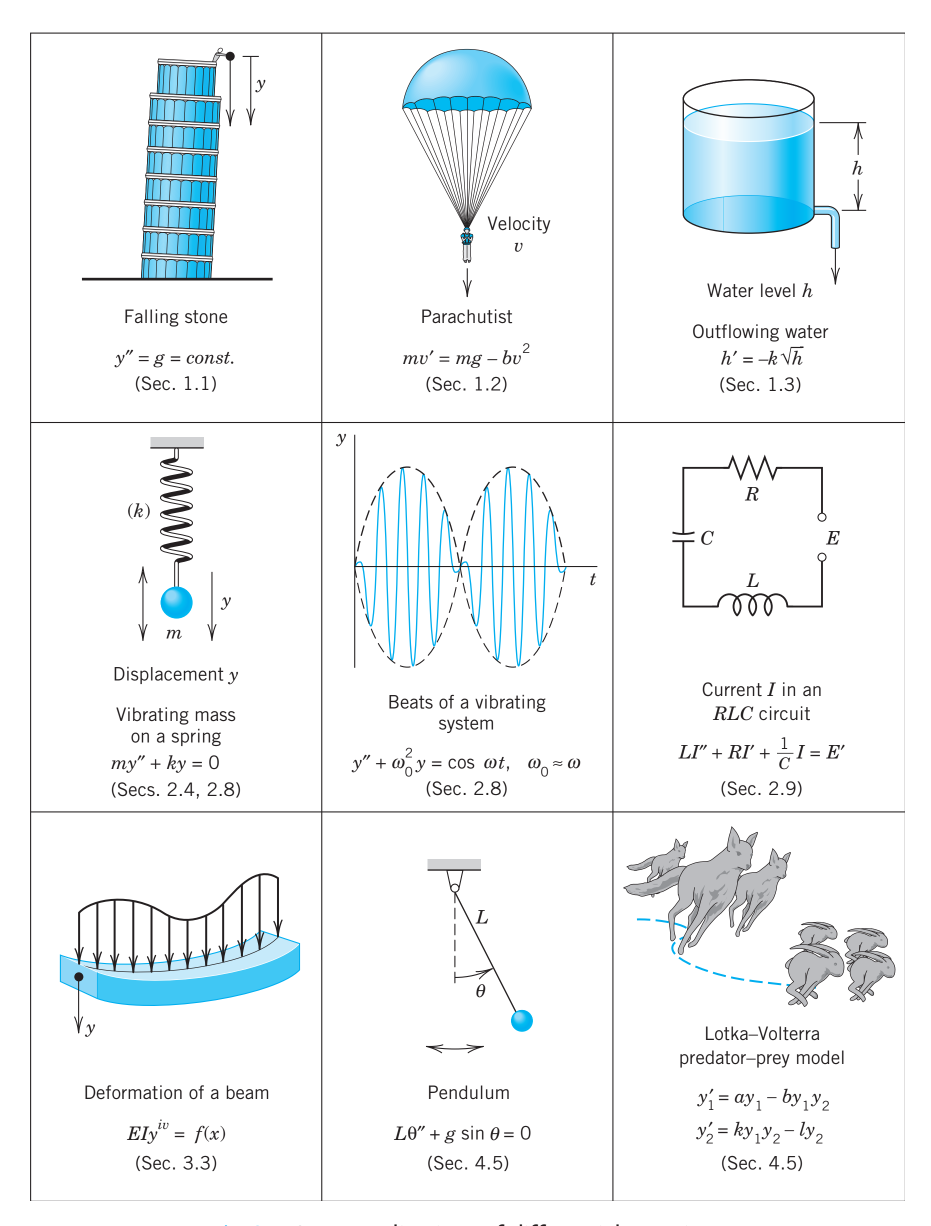


Fig. 2. Some applications of differential equations

An **ordinary differential equation (ODE)** is an equation that contains one or several derivatives of an unknown function, which we usually call $y(x)$ (or sometimes $y(t)$ if the independent variable is time t). The equation may also contain y itself, known functions of x (or t), and constants. For example,

$$y' = \cos x \tag{1}$$

$$y'' + 9y = e^{-2x} \tag{2}$$

$$y'y''' - \tfrac{3}{2}y'^2 = 0 \tag{3}$$

are ordinary differential equations (ODEs). Here, as in calculus, y' denotes dy/dx, $y'' = d^2y/dx^2$, etc. The term *ordinary* distinguishes them from *partial differential equations* (PDEs), which involve partial derivatives of an unknown function of *two or more* variables. For instance, a PDE with unknown function u of two variables x and y is

$$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0.$$

PDEs have important engineering applications, but they are more complicated than ODEs; they will be considered in Chap. 12.

An ODE is said to be of **order** n if the nth derivative of the unknown function y is the highest derivative of y in the equation. The concept of order gives a useful classification into ODEs of first order, second order, and so on. Thus, (1) is of first order, (2) of second order, and (3) of third order.

In this chapter we shall consider **first-order ODEs**. Such equations contain only the first derivative y' and may contain y and any given functions of x. Hence we can write them as

(4) $$F(x, y, y') = 0$$

or often in the form

$$y' = f(x, y).$$

This is called the *explicit form*, in contrast to the *implicit form* (4). For instance, the implicit ODE $x^{-3}y' - 4y^2 = 0$ (where $x \neq 0$) can be written explicitly as $y' = 4x^3y^2$.

Concept of Solution

A function

$$y = h(x)$$

is called a **solution** of a given ODE (4) on some open interval $a < x < b$ if $h(x)$ is defined and differentiable throughout the interval and is such that the equation becomes an identity if y and y' are replaced with h and h', respectively. The curve (the graph) of h is called a **solution curve**.

Here, **open interval** $a < x < b$ means that the endpoints a and b are not regarded as points belonging to the interval. Also, $a < x < b$ includes *infinite intervals* $-\infty < x < b$, $a < x < \infty$, $-\infty < x < \infty$ (the real line) as special cases.

EXAMPLE 1 Verification of Solution

Verify that $y = c/x$ (c an arbitrary constant) is a solution of the ODE $xy' = -y$ for all $x \neq 0$. Indeed, differentiate $y = c/x$ to get $y' = -c/x^2$. Multiply this by x, obtaining $xy' = -c/x$; thus, $xy' = -y$, the given ODE. ■

EXAMPLE 2 **Solution by Calculus. Solution Curves**

The ODE $y' = dy/dx = \cos x$ can be solved directly by integration on both sides. Indeed, using calculus, we obtain $y = \int \cos x\, dx = \sin x + c$, where c is an arbitrary constant. This is a ***family of solutions***. Each value of c, for instance, 2.75 or 0 or -8, gives one of these curves. Figure 3 shows some of them, for $c = -3, -2, -1, 0, 1, 2, 3, 4$. ■

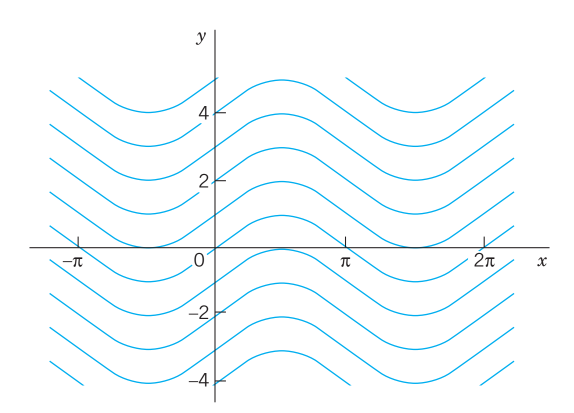


Fig. 3. Solutions $y = \sin x + c$ of the ODE $y' = \cos x$

EXAMPLE 3 **(A) Exponential Growth. (B) Exponential Decay**

From calculus we know that $y = ce^{0.2t}$ has the derivative

$$y' = \frac{dy}{dt} = 0.2e^{0.2t} = 0.2y.$$

Hence y is a solution of $y' = 0.2y$ (Fig. 4A). This ODE is of the form $y' = ky$. With positive-constant k it can model exponential growth, for instance, of colonies of bacteria or populations of animals. It also applies to humans for small populations in a large country (e.g., the United States in early times) and is then known as **Malthus's law**.[1] We shall say more about this topic in Sec. 1.5.

(B) Similarly, $y' = -0.2$ (with a minus on the right) has the solution $y = ce^{-0.2t}$, (Fig. 4B) modeling **exponential decay**, as, for instance, of a radioactive substance (see Example 5). ■

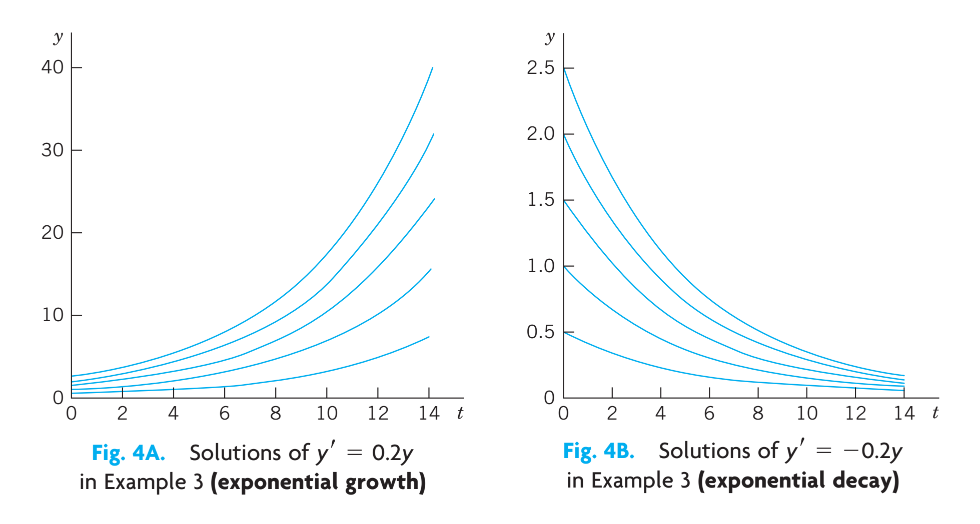


Fig. 4A. Solutions of $y' = 0.2y$ in Example 3 **(exponential growth)**

Fig. 4B. Solutions of $y' = -0.2y$ in Example 3 **(exponential decay)**

[1]Named after the English pioneer in classic economics, THOMAS ROBERT MALTHUS (1766–1834).

We see that each ODE in these examples has a solution that contains an arbitrary constant c. Such a solution containing an arbitrary constant c is called a **general solution** of the ODE.

(We shall see that c is sometimes not completely arbitrary but must be restricted to some interval to avoid complex expressions in the solution.)

We shall develop methods that will give general solutions *uniquely* (perhaps except for notation). Hence we shall say ***the*** general solution of a given ODE (instead of ***a*** general solution).

Geometrically, the general solution of an ODE is a family of infinitely many solution curves, one for each value of the constant c. If we choose a specific c (e.g., $c = 6.45$ or 0 or -2.01) we obtain what is called a **particular solution** of the ODE. A particular solution does not contain any arbitrary constants.

In most cases, general solutions exist, and every solution not containing an arbitrary constant is obtained as a particular solution by assigning a suitable value to c. Exceptions to these rules occur but are of minor interest in applications; see Prob. 16 in Problem Set 1.1.

Initial Value Problem

In most cases the unique solution of a given problem, hence a particular solution, is obtained from a general solution by an **initial condition** $y(x_0) = y_0$, with given values x_0 and y_0, that is used to determine a value of the arbitrary constant c. Geometrically this condition means that the solution curve should pass through the point (x_0, y_0) in the xy-plane. An ODE, together with an initial condition, is called an **initial value problem**. Thus, if the ODE is explicit, $y' = f(x, y)$, the initial value problem is of the form

$$y' = f(x, y), \qquad y(x_0) = y_0. \tag{5}$$

EXAMPLE 4 Initial Value Problem

Solve the initial value problem

$$y' = \frac{dy}{dx} = 3y, \qquad y(0) = 5.7.$$

Solution. The general solution is $y(x) = ce^{3x}$; see Example 3. From this solution and the initial condition we obtain $y(0) = ce^0 = c = 5.7$. Hence the initial value problem has the solution $y(x) = 5.7e^{3x}$. This is a particular solution. ■

More on Modeling

The general importance of modeling to the engineer and physicist was emphasized at the beginning of this section. We shall now consider a basic physical problem that will show the details of the typical steps of modeling. Step 1: the transition from the physical situation (the physical system) to its mathematical formulation (its mathematical model); Step 2: the solution by a mathematical method; and Step 3: the physical interpretation of the result. This may be the easiest way to obtain a first idea of the nature and purpose of differential equations and their applications. Realize at the outset that your ***computer*** (your ***CAS***) may perhaps give you a hand in Step 2, but Steps 1 and 3 are basically your work.

And Step 2 requires a solid knowledge and good understanding of solution methods available to you—*you* have to choose the method for your work by hand or by the computer. Keep this in mind, and always check computer results for errors (which may arise, for instance, from false inputs).

EXAMPLE 5 Radioactivity. Exponential Decay

Given an amount of a radioactive substance, say, 0.5 g (gram), find the amount present at any later time.

Physical Information. Experiments show that at each instant a radioactive substance decomposes—and is thus decaying in time—proportional to the amount of substance present.

Step 1. Setting up a mathematical model of the physical process. Denote by $y(t)$ the amount of substance still present at any time t. By the physical law, the time rate of change $y'(t) = dy/dt$ is proportional to $y(t)$. This gives the ***first-order ODE***

$$\frac{dy}{dt} = -ky \tag{6}$$

where the constant k is positive, so that, because of the minus, we do get decay (as in [B] of Example 3). The value of k is known from experiments for various radioactive substances (e.g., $k = 1.4 \cdot 10^{-11}\ \text{sec}^{-1}$, approximately, for radium $^{226}_{88}\text{Ra}$).

Now the given initial amount is 0.5 g, and we can call the corresponding instant $t = 0$. Then we have the ***initial condition*** $y(0) = 0.5$. This is the instant at which our observation of the process begins. It motivates the term *initial condition* (which, however, is also used when the independent variable is not time or when we choose a t other than $t = 0$). Hence the mathematical model of the physical process is the ***initial value problem***

$$\frac{dy}{dt} = -ky, \qquad y(0) = 0.5. \tag{7}$$

Step 2. Mathematical solution. As in (B) of Example 3 we conclude that the ODE (6) models exponential decay and has the general solution (with arbitrary constant c but definite given k)

$$y(t) = ce^{-kt}. \tag{8}$$

We now determine c by using the initial condition. Since $y(0) = c$ from (8), this gives $y(0) = c = 0.5$. Hence the particular solution governing our process is (cf. Fig. 5)

$$y(t) = 0.5e^{-kt} \qquad (k > 0). \tag{9}$$

Always check your result—it may involve human or computer errors! Verify by differentiation (chain rule!) that your solution (9) satisfies (7) as well as $y(0) = 0.5$:

$$\frac{dy}{dt} = -0.5ke^{-kt} = -k \cdot 0.5e^{-kt} = -ky, \qquad y(0) = 0.5e^0 = 0.5.$$

Step 3. Interpretation of result. Formula (9) gives the amount of radioactive substance at time t. It starts from the correct initial amount and decreases with time because k is positive. The limit of y as $t \to \infty$ is zero. ■

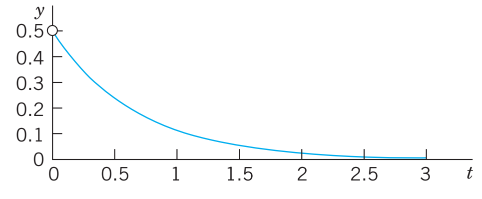


Fig. 5. Radioactivity (Exponential decay, $y = 0.5e^{-kt}$, with $k = 1.5$ as an example)

PROBLEM SET 1.1

1–8 CALCULUS

Solve the ODE by integration or by remembering a differentiation formula.

1. $y' + 2 \sin 2\pi x = 0$

2. $y' + xe^{-x^2/2} = 0$

3. $y' = y$

4. $y' = -1.5y$

5. $y' = 4e^{-x} \cos x$

6. $y'' = -y$

7. $y' = \cosh 5.13x$

8. $y''' = e^{-0.2x}$

9–15 VERIFICATION. INITIAL VALUE PROBLEM (IVP)

(a) Verify that y is a solution of the ODE. **(b)** Determine from y the particular solution of the IVP. **(c)** Graph the solution of the IVP.

9. $y' + 4y = 1.4, \quad y = ce^{-4x} + 0.35, \quad y(0) = 2$

10. $y' + 5xy = 0, \quad y = ce^{-2.5x^2}, \quad y(0) = \pi$

11. $y' = y + e^x, \quad y = (x + c)e^x, \quad y(0) = \frac{1}{2}$

12. $yy' = 4x, \quad y^2 - 4x^2 = c \ (y > 0), \quad y(1) = 4$

13. $y' = y - y^2, \quad y = \dfrac{1}{1 + ce^{-x}}, \quad y(0) = 0.25$

14. $y' \tan x = 2y - 8, \quad y = c \sin^2 x + 4, \quad y(\frac{1}{2}\pi) = 0$

15. Find two constant solutions of the ODE in Prob. 13 by inspection.

16. Singular solution. An ODE may sometimes have an additional solution that cannot be obtained from the general solution and is then called a *singular solution.* The ODE $y'^2 - xy' + y = 0$ is of this kind. Show by differentiation and substitution that it has the general solution $y = cx - c^2$ and the singular solution $y = x^2/4$. Explain Fig. 6.

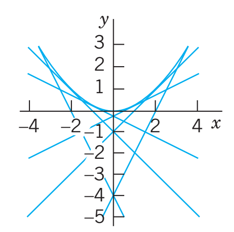


Fig. 6. Particular solutions and singular solution in Problem 16

17–20 MODELING, APPLICATIONS

These problems will give you a first impression of modeling. Many more problems on modeling follow throughout this chapter.

17. Half-life. The *half-life* measures exponential decay. It is the time in which half of the given amount of radioactive substance will disappear. What is the half-life of $^{226}_{88}\text{Ra}$ (in years) in Example 5?

18. Half-life. Radium $^{224}_{88}\text{Ra}$ has a half-life of about 3.6 days.

(a) Given 1 gram, how much will still be present after 1 day?

(b) After 1 year?

19. Free fall. In dropping a stone or an iron ball, air resistance is practically negligible. Experiments show that the acceleration of the motion is constant (equal to $g = 9.80 \text{ m/sec}^2 = 32 \text{ ft/sec}^2$, called the **acceleration of gravity**). Model this as an ODE for $y(t)$, the distance fallen as a function of time t. If the motion starts at time $t = 0$ from rest (i.e., with velocity $v = y' = 0$), show that you obtain the familiar law of free fall

$$y = \tfrac{1}{2}gt^2.$$

20. Exponential decay. Subsonic flight. The efficiency of the engines of subsonic airplanes depends on air pressure and is usually maximum near 35,000 ft. Find the air pressure $y(x)$ at this height. *Physical information.* The rate of change $y'(x)$ is proportional to the pressure. At 18,000 ft it is half its value $y_0 = y(0)$ at sea level. *Hint.* Remember from calculus that if $y = e^{kx}$, then $y' = ke^{kx} = ky$. Can you see without calculation that the answer should be close to $y_0/4$?

1.2 Geometric Meaning of $y' = f(x, y)$. Direction Fields, Euler's Method

A first-order ODE

$$y' = f(x, y) \tag{1}$$

has a simple geometric interpretation. From calculus you know that the derivative $y'(x)$ of $y(x)$ is the slope of $y(x)$. Hence a solution curve of (1) that passes through a point (x_0, y_0) must have, at that point, the slope $y'(x_0)$ equal to the value of f at that point; that is,

$$y'(x_0) = f(x_0, y_0).$$

Using this fact, we can develop graphic or numeric methods for obtaining approximate solutions of ODEs (1). This will lead to a better conceptual understanding of an ODE (1). Moreover, such methods are of practical importance since many ODEs have complicated solution formulas or no solution formulas at all, whereby numeric methods are needed.

Graphic Method of Direction Fields. Practical Example Illustrated in Fig. 7. We can show directions of solution curves of a given ODE (1) by drawing short straight-line segments (lineal elements) in the xy-plane. This gives a **direction field** (or *slope field*) into which you can then fit (approximate) solution curves. This may reveal typical properties of the whole family of solutions.

Figure 7 shows a direction field for the ODE

$$y' = y + x \tag{2}$$

obtained by a CAS (Computer Algebra System) and some approximate solution curves fitted in.

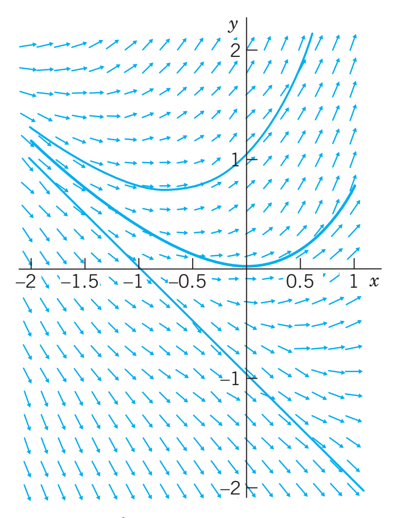


Fig. 7. Direction field of $y' = y + x$, with three approximate solution curves passing through (0, 1), (0, 0), (0, −1), respectively

If you have no CAS, first draw a few *level curves* $f(x, y) = \text{const}$ of $f(x, y)$, then parallel lineal elements along each such curve (which is also called an **isocline**, meaning a curve of equal inclination), and finally draw approximation curves fit to the lineal elements.

We shall now illustrate how numeric methods work by applying the simplest numeric method, that is Euler's method, to an initial value problem involving ODE (2). First we give a brief description of Euler's method.

Numeric Method by Euler

Given an ODE (1) and an initial value $y(x_0) = y_0$, **Euler's method** yields approximate solution values at equidistant x-values $x_0, x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots$, namely,

$$y_1 = y_0 + hf(x_0, y_0) \qquad \text{(Fig. 8)}$$

$$y_2 = y_1 + hf(x_1, y_1), \qquad \text{etc.}$$

In general,

$$y_n = y_{n-1} + hf(x_{n-1}, y_{n-1})$$

where the step h equals, e.g., 0.1 or 0.2 (as in Table 1.1) or a smaller value for greater accuracy.

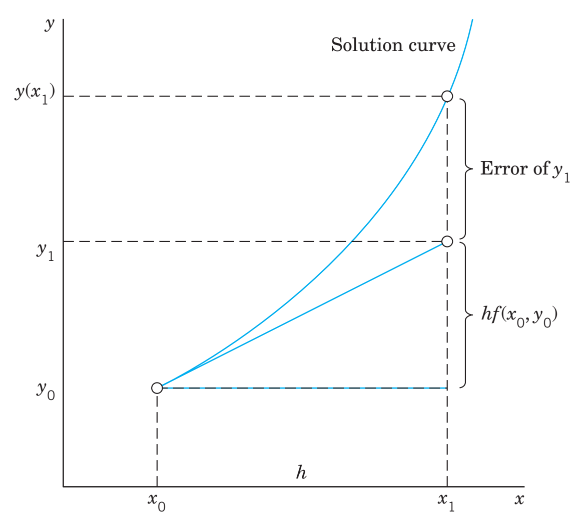


Fig. 8. First Euler step, showing a solution curve, its tangent at (x_0, y_0), step h and increment $hf(x_0, y_0)$ in the formula for y_1

Table 1.1 shows the computation of $n = 5$ steps with step $h = 0.2$ for the ODE (2) and initial condition $y(0) = 0$, corresponding to the middle curve in the direction field. We shall solve the ODE exactly in Sec. 1.5. For the time being, verify that the initial value problem has the solution $y = e^x - x - 1$. The solution curve and the values in Table 1.1 are shown in Fig. 9. These values are rather inaccurate. The errors $y(x_n) - y_n$ are shown in Table 1.1 as well as in Fig. 9. Decreasing h would improve the values, but would soon require an impractical amount of computation. Much better methods of a similar nature will be discussed in Sec. 21.1.

Table 1.1. **Euler method for $y' = y + x, y(0) = 0$ for $x = 0, \cdots, 1.0$ with step $h = 0.2$**

n	x_n	y_n	$y(x_n)$	Error
0	0.0	0.000	0.000	0.000
1	0.2	0.000	0.021	0.021
2	0.4	0.04	0.092	0.052
3	0.6	0.128	0.222	0.094
4	0.8	0.274	0.426	0.152
5	1.0	0.488	0.718	0.230

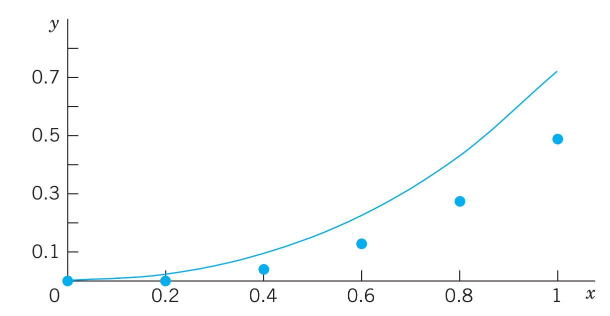


Fig. 9. Euler method: Approximate values in Table 1.1 and solution curve

PROBLEM SET 1.2

1–8 DIRECTION FIELDS, SOLUTION CURVES

Graph a direction field (by a CAS or by hand). In the field graph several solution curves by hand, particularly those passing through the given points (x, y).

1. $y' = 1 + y^2$, $(\frac{1}{4}\pi, 1)$

2. $yy' + 4x = 0$, $(1, 1), (0, 2)$

3. $y' = 1 - y^2$, $(0, 0), (2, \frac{1}{2})$

4. $y' = 2y - y^2$, $(0, 0), (0, 1), (0, 2), (0, 3)$

5. $y' = x - 1/y$, $(1, \frac{1}{2})$

6. $y' = \sin^2 y$, $(0, -0.4), (0, 1)$

7. $y' = e^{y/x}$, $(2, 2), (3, 3)$

8. $y' = -2xy$, $(0, \frac{1}{2}), (0, 1), (0, 2)$

9–10 ACCURACY OF DIRECTION FIELDS

Direction fields are very useful because they can give you an impression of all solutions without solving the ODE, which may be difficult or even impossible. To get a feel for the accuracy of the method, graph a field, sketch solution curves in it, and compare them with the exact solutions.

9. $y' = \cos \pi x$

10. $y' = -5y^{1/2}$ (Sol. $\sqrt{y} + \frac{5}{2}x = c$)

11. Autonomous ODE. This means an ODE not showing x (the independent variable) *explicitly*. (The ODEs in Probs. 6 and 10 are autonomous.) What will the level curves $f(x, y) = \text{const}$ (also called *isoclines* = curves of equal inclination) of an autonomous ODE look like? Give reason.

12–15 MOTIONS

Model the motion of a body B on a straight line with velocity as given, $y(t)$ being the distance of B from a point $y = 0$ at time t. Graph a direction field of the model (the ODE). In the field sketch the solution curve satisfying the given initial condition.

12. Product of velocity times distance constant, equal to 2, $y(0) = 2$.

13. Distance = Velocity × Time, $y(1) = 1$

14. Square of the distance plus square of the velocity equal to 1, initial distance $1/\sqrt{2}$

15. Parachutist. Two forces act on a parachutist, the attraction by the earth mg (m = mass of person plus equipment, $g = 9.8 \text{ m/sec}^2$ the acceleration of gravity) and the air resistance, assumed to be proportional to the square of the velocity $v(t)$. Using **Newton's second law** of motion (mass × acceleration = resultant of the forces), set up a model (an ODE for $v(t)$). Graph a direction field (choosing m and the constant of proportionality equal to 1). Assume that the parachute opens when $v = 10$ m/sec. Graph the corresponding solution in the field. What is the limiting velocity? Would the parachute still be sufficient if the air resistance were only proportional to $v(t)$?

16. CAS PROJECT. Direction Fields. Discuss direction fields as follows.

(a) Graph portions of the direction field of the ODE (2) (see Fig. 7), for instance, $-5 \leqq x \leqq 2$, $-1 \leqq y \leqq 5$. Explain what you have gained by this enlargement of the portion of the field.

(b) Using implicit differentiation, find an ODE with the general solution $x^2 + 9y^2 = c$ $(y > 0)$. Graph its direction field. Does the field give the impression that the solution curves may be semi-ellipses? Can you do similar work for circles? Hyperbolas? Parabolas? Other curves?

(c) Make a conjecture about the solutions of $y' = -x/y$ from the direction field.

(d) Graph the direction field of $y' = -\frac{1}{2}y$ and some solutions of your choice. How do they behave? Why do they decrease for $y > 0$?

17–20 EULER'S METHOD

This is the simplest method to explain numerically solving an ODE, more precisely, an initial value problem (IVP). (More accurate methods based on the same principle are explained in Sec. 21.1.) Using the method, to get a feel for numerics as well as for the nature of IVPs, solve the IVP numerically with a PC or a calculator, 10 steps. Graph the computed values and the solution curve on the same coordinate axes.

17. $y' = y, \quad y(0) = 1, \quad h = 0.1$

18. $y' = y, \quad y(0) = 1, \quad h = 0.01$

19. $y' = (y - x)^2, \quad y(0) = 0, \quad h = 0.1$
Sol. $y = x - \tanh x$

20. $y' = -5x^4y^2, \quad y(0) = 1, \quad h = 0.2$
Sol. $y = 1/(1 + x)^5$

1.3 Separable ODEs. Modeling

Many practically useful ODEs can be reduced to the form

$$g(y)\, y' = f(x) \tag{1}$$

by purely algebraic manipulations. Then we can integrate on both sides with respect to x, obtaining

$$\int g(y)\, y'\, dx = \int f(x)\, dx + c. \tag{2}$$

On the left we can switch to y as the variable of integration. By calculus, $y'\,dx = dy$, so that

$$\int g(y)\, dy = \int f(x)\, dx + c. \tag{3}$$

If f and g are continuous functions, the integrals in (3) exist, and by evaluating them we obtain a general solution of (1). This method of solving ODEs is called the **method of separating variables**, and (1) is called a **separable equation**, because in (3) the variables are now separated: x appears only on the right and y only on the left.

EXAMPLE 1 Separable ODE

The ODE $y' = 1 + y^2$ is separable because it can be written

$$\frac{dy}{1 + y^2} = dx. \qquad \text{By integration,} \qquad \arctan y = x + c \qquad \text{or} \qquad y = \tan(x + c).$$

It is very important to introduce the constant of integration immediately when the integration is performed. If we wrote $\arctan y = x$, then $y = \tan x$, and *then* introduced c, we would have obtained $y = \tan x + c$, which is not a solution (when $c \neq 0$). Verify this. ■

EXAMPLE 2 Separable ODE

The ODE $y' = (x + 1)e^{-x}y^2$ is separable; we obtain $y^{-2}\,dy = (x + 1)e^{-x}\,dx$.

By integration, $\quad -y^{-1} = -(x + 2)e^{-x} + c, \qquad y = \dfrac{1}{(x + 2)e^{-x} - c}.$ ■

EXAMPLE 3 Initial Value Problem (IVP). Bell-Shaped Curve

Solve $y' = -2xy$, $y(0) = 1.8$.

Solution. By separation and integration,

$$\frac{dy}{y} = -2x\,dx, \qquad \ln y = -x^2 + \tilde{c}, \qquad y = ce^{-x^2}.$$

This is the general solution. From it and the initial condition, $y(0) = ce^0 = c = 1.8$. Hence the IVP has the solution $y = 1.8e^{-x^2}$. This is a particular solution, representing a bell-shaped curve (Fig. 10). ■

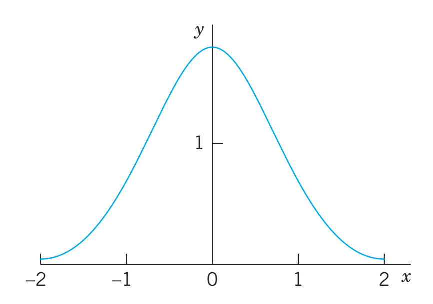


Fig. 10. Solution in Example 3 (bell-shaped curve)

Modeling

The importance of modeling was emphasized in Sec. 1.1, and separable equations yield various useful models. Let us discuss this in terms of some typical examples.

EXAMPLE 4 Radiocarbon Dating[2]

In September 1991 the famous Iceman (Oetzi), a mummy from the Neolithic period of the Stone Age found in the ice of the Oetztal Alps (hence the name "Oetzi") in Southern Tyrolia near the Austrian–Italian border, caused a scientific sensation. When did Oetzi approximately live and die if the ratio of carbon ${}^{14}_{6}\mathrm{C}$ to carbon ${}^{12}_{6}\mathrm{C}$ in this mummy is 52.5% of that of a living organism?

Physical Information. In the atmosphere and in living organisms, the ratio of radioactive carbon ${}^{14}_{6}\mathrm{C}$ (made radioactive by cosmic rays) to ordinary carbon ${}^{12}_{6}\mathrm{C}$ is constant. When an organism dies, its absorption of ${}^{14}_{6}\mathrm{C}$ by breathing and eating terminates. Hence one can estimate the age of a fossil by comparing the radioactive carbon ratio in the fossil with that in the atmosphere. To do this, one needs to know the half-life of ${}^{14}_{6}\mathrm{C}$, which is 5715 years (*CRC Handbook of Chemistry and Physics,* 83rd ed., Boca Raton: CRC Press, 2002, page 11–52, line 9).

Solution. ***Modeling.*** Radioactive decay is governed by the ODE $y' = ky$ (see Sec. 1.1, Example 5). By separation and integration (where t is time and y_0 is the initial ratio of ${}^{14}_{6}\mathrm{C}$ to ${}^{12}_{6}\mathrm{C}$)

$$\frac{dy}{y} = k\,dt, \qquad \ln|y| = kt + c, \qquad y = y_0 e^{kt} \qquad (y_0 = e^c).$$

---

[2]Method by WILLARD FRANK LIBBY (1908–1980), American chemist, who was awarded for this work the 1960 Nobel Prize in chemistry.

Next we use the half-life $H = 5715$ to determine k. When $t = H$, half of the original substance is still present. Thus,

$$y_0 e^{kH} = 0.5 y_0, \qquad e^{kH} = 0.5, \qquad k = \frac{\ln 0.5}{H} = -\frac{0.693}{5715} = -0.0001213.$$

Finally, we use the ratio 52.5% for determining the time t when Oetzi died (actually, was killed),

$$e^{kt} = e^{-0.0001213t} = 0.525, \qquad t = \frac{\ln 0.525}{-0.0001213} = 5312. \qquad \textit{Answer:} \qquad \text{About 5300 years ago.}$$

Other methods show that radiocarbon dating values are usually too small. According to recent research, this is due to a variation in that carbon ratio because of industrial pollution and other factors, such as nuclear testing. ■

EXAMPLE 5 Mixing Problem

Mixing problems occur quite frequently in chemical industry. We explain here how to solve the basic model involving a single tank. The tank in Fig. 11 contains 1000 gal of water in which initially 100 lb of salt is dissolved. Brine runs in at a rate of 10 gal/min, and each gallon contains 5 lb of dissoved salt. The mixture in the tank is kept uniform by stirring. Brine runs out at 10 gal/min. Find the amount of salt in the tank at any time t.

Solution. ***Step 1. Setting up a model.*** Let $y(t)$ denote the amount of salt in the tank at time t. Its time rate of change is

$$y' = \text{Salt inflow rate} - \text{Salt outflow rate} \qquad \textbf{Balance law.}$$

5 lb times 10 gal gives an inflow of 50 lb of salt. Now, the outflow is 10 gal of brine. This is $10/1000 = 0.01$ $(= 1\%)$ of the total brine content in the tank, hence 0.01 of the salt content $y(t)$, that is, 0.01 $y(t)$. Thus the model is the ODE

$$y' = 50 - 0.01y = -0.01(y - 5000). \tag{4}$$

Step 2. Solution of the model. The ODE (4) is separable. Separation, integration, and taking exponents on both sides gives

$$\frac{dy}{y - 5000} = -0.01\, dt, \qquad \ln|y - 5000| = -0.01t + c^*, \qquad y - 5000 = ce^{-0.01t}.$$

Initially the tank contains 100 lb of salt. Hence $y(0) = 100$ is the initial condition that will give the unique solution. Substituting $y = 100$ and $t = 0$ in the last equation gives $100 - 5000 = ce^0 = c$. Hence $c = -4900$. Hence the amount of salt in the tank at time t is

$$y(t) = 5000 - 4900e^{-0.01t}. \tag{5}$$

This function shows an exponential approach to the limit 5000 lb; see Fig. 11. Can you explain physically that $y(t)$ should increase with time? That its limit is 5000 lb? Can you see the limit directly from the ODE?

The model discussed becomes more realistic in problems on pollutants in lakes (see Problem Set 1.5, Prob. 35) or drugs in organs. These types of problems are more difficult because the mixing may be imperfect and the flow rates (in and out) may be different and known only very roughly. ■

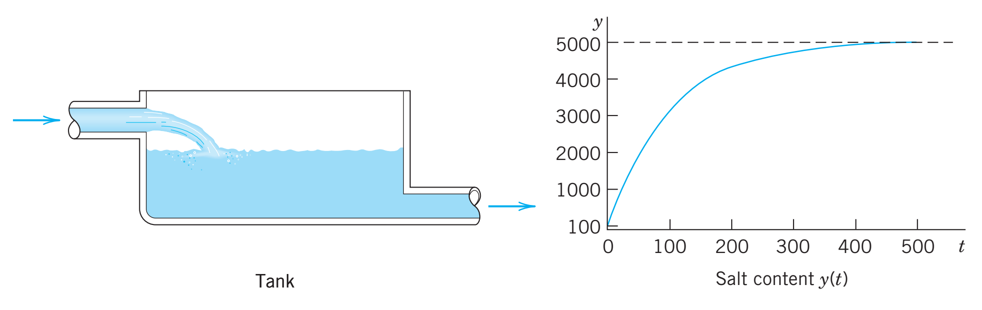


Fig. 11. Mixing problem in Example 5

EXAMPLE 6 **Heating an Office Building (Newton's Law of Cooling[3])**

Suppose that in winter the daytime temperature in a certain office building is maintained at 70°F. The heating is shut off at 10 P.M. and turned on again at 6 A.M. On a certain day the temperature inside the building at 2 A.M. was found to be 65°F. The outside temperature was 50°F at 10 P.M. and had dropped to 40°F by 6 A.M. What was the temperature inside the building when the heat was turned on at 6 A.M.?

Physical information. Experiments show that the time rate of change of the temperature T of a body B (which conducts heat well, for example, as a copper ball does) is proportional to the difference between T and the temperature of the surrounding medium **(Newton's law of cooling).**

Solution. ***Step 1. Setting up a model.*** Let $T(t)$ be the temperature inside the building and T_A the outside temperature (assumed to be constant in Newton's law). Then by Newton's law,

$$\frac{dT}{dt} = k(T - T_A). \tag{6}$$

Such experimental laws are derived under idealized assumptions that rarely hold exactly. However, even if a model seems to fit the reality only poorly (as in the present case), it may still give valuable qualitative information. To see how good a model is, the engineer will collect experimental data and compare them with calculations from the model.

Step 2. General solution. We cannot solve (6) because we do not know T_A, just that it varied between 50°F and 40°F, so we follow the ***Golden Rule:*** *If you cannot solve your problem, try to solve a simpler one.* We solve (6) with the unknown function T_A replaced with the average of the two known values, or 45°F. For physical reasons we may expect that this will give us a reasonable approximate value of T in the building at 6 A.M.

For constant $T_A = 45$ (or any other *constant* value) the ODE (6) is separable. Separation, integration, and taking exponents gives the general solution

$$\frac{dT}{T-45} = k\,dt, \qquad \ln|T - 45| = kt + c^*, \qquad T(t) = 45 + ce^{kt} \qquad (c = e^{c^*}).$$

Step 3. Particular solution. We choose 10 P.M. to be $t = 0$. Then the given initial condition is $T(0) = 70$ and yields a particular solution, call it T_p. By substitution,

$$T(0) = 45 + ce^0 = 70, \qquad c = 70 - 45 = 25, \qquad T_p(t) = 45 + 25e^{kt}.$$

Step 4. Determination of k. We use $T(4) = 65$, where $t = 4$ is 2 A.M. Solving algebraically for k and inserting k into $T_p(t)$ gives (Fig. 12)

$$T_p(4) = 45 + 25e^{4k} = 65, \qquad e^{4k} = 0.8, \qquad k = \tfrac{1}{4}\ln 0.8 = -0.056, \qquad T_p(t) = 45 + 25e^{-0.056t}.$$

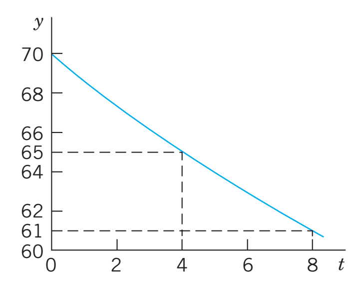


Fig. 12. Particular solution (temperature) in Example 6

[3] Sir ISAAC NEWTON (1642–1727), great English physicist and mathematician, became a professor at Cambridge in 1669 and Master of the Mint in 1699. He and the German mathematician and philosopher GOTTFRIED WILHELM LEIBNIZ (1646–1716) invented (independently) the differential and integral calculus. Newton discovered many basic physical laws and created the method of investigating physical problems by means of calculus. His *Philosophiae naturalis principia mathematica* (*Mathematical Principles of Natural Philosophy,* 1687) contains the development of classical mechanics. His work is of greatest importance to both mathematics and physics.

Step 5. Answer and interpretation. 6 A.M. is $t = 8$ (namely, 8 hours after 10 P.M.), and

$$T_p(8) = 45 + 25e^{-0.056 \cdot 8} = 61[°F].$$

Hence the temperature in the building dropped 9°F, a result that looks reasonable. ■

EXAMPLE 7 Leaking Tank. Outflow of Water Through a Hole (Torricelli's Law)

This is another prototype engineering problem that leads to an ODE. It concerns the outflow of water from a cylindrical tank with a hole at the bottom (Fig. 13). You are asked to find the height of the water in the tank at any time if the tank has diameter 2 m, the hole has diameter 1 cm, and the initial height of the water when the hole is opened is 2.25 m. When will the tank be empty?

Physical information. Under the influence of gravity the outflowing water has velocity

$$v(t) = 0.600\sqrt{2gh(t)} \qquad \textbf{(Torricelli's law[4])}, \tag{7}$$

where $h(t)$ is the height of the water above the hole at time t, and $g = 980 \text{ cm/sec}^2 = 32.17 \text{ ft/sec}^2$ is the acceleration of gravity at the surface of the earth.

Solution. ***Step 1. Setting up the model.*** To get an equation, we relate the decrease in water level $h(t)$ to the outflow. The volume ΔV of the outflow during a short time Δt is

$$\Delta V = Av\,\Delta t \qquad (A = \text{Area of hole}).$$

ΔV must equal the change ΔV^* of the volume of the water in the tank. Now

$$\Delta V^* = -B\,\Delta h \qquad (B = \text{Cross-sectional area of tank})$$

where $\Delta h\ (> 0)$ is the decrease of the height $h(t)$ of the water. The minus sign appears because the volume of the water in the tank decreases. Equating ΔV and ΔV^* gives

$$-B\,\Delta h = Av\,\Delta t.$$

We now express v according to Torricelli's law and then let Δt (the length of the time interval considered) approach 0—this is a ***standard way*** of obtaining an ODE as a model. That is, we have

$$\frac{\Delta h}{\Delta t} = -\frac{A}{B}v = -\frac{A}{B}0.600\sqrt{2gh(t)}$$

and by letting $\Delta t \to 0$ we obtain the ODE

$$\frac{dh}{dt} = -26.56\frac{A}{B}\sqrt{h},$$

where $26.56 = 0.600\sqrt{2 \cdot 980}$. This is our model, a first-order ODE.

Step 2. General solution. Our ODE is separable. A/B is constant. Separation and integration gives

$$\frac{dh}{\sqrt{h}} = -26.56\frac{A}{B}dt \qquad \text{and} \qquad 2\sqrt{h} = c^* - 26.56\frac{A}{B}t.$$

Dividing by 2 and squaring gives $h = (c - 13.28At/B)^2$. Inserting $13.28A/B = 13.28 \cdot 0.5^2\pi/100^2\pi = 0.000332$ yields the general solution

$$h(t) = (c - 0.000332t)^2.$$

[4]EVANGELISTA TORRICELLI (1608–1647), Italian physicist, pupil and successor of GALILEO GALILEI (1564–1642) at Florence. The "contraction factor" 0.600 was introduced by J. C. BORDA in 1766 because the stream has a smaller cross section than the area of the hole.

Step 3. Particular solution. The initial height (the initial condition) is $h(0) = 225$ cm. Substitution of $t = 0$ and $h = 225$ gives from the general solution $c^2 = 225$, $c = 15.00$ and thus the particular solution (Fig. 13)

$$h_p(t) = (15.00 - 0.000332t)^2.$$

Step 4. Tank empty. $h_p(t) = 0$ if $t = 15.00/0.000332 = 45{,}181$ [sec] $= 12.6$ [hours].

Here you see distinctly the ***importance of the choice of units***—we have been working with the cgs system, in which time is measured in seconds! We used $g = 980\ \text{cm/sec}^2$.

Step 5. Checking. Check the result. ■

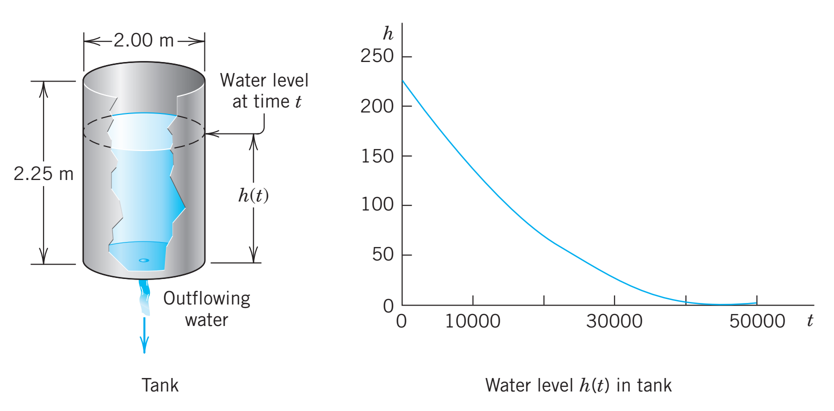


Fig. 13. Example 7. Outflow from a cylindrical tank ("leaking tank"). Torricelli's law

Extended Method: Reduction to Separable Form

Certain nonseparable ODEs can be made separable by transformations that introduce for y a new unknown function. We discuss this technique for a class of ODEs of practical importance, namely, for equations

$$y' = f\left(\frac{y}{x}\right). \tag{8}$$

Here, f is any (differentiable) function of y/x, such as $\sin(y/x)$, $(y/x)^4$, and so on. (Such an ODE is sometimes called a *homogeneous ODE,* a term we shall not use but reserve for a more important purpose in Sec. 1.5.)

The form of such an ODE suggests that we set $y/x = u$; thus,

(9) $\quad y = ux \quad$ and by product differentiation $\quad y' = u'x + u.$

Substitution into $y' = f(y/x)$ then gives $u'x + u = f(u)$ or $u'x = f(u) - u$. We see that if $f(u) - u \neq 0$, this can be separated:

$$\frac{du}{f(u) - u} = \frac{dx}{x}. \tag{10}$$

EXAMPLE 8 Reduction to Separable Form

Solve

$$2xyy' = y^2 - x^2.$$

Solution. To get the usual explicit form, divide the given equation by $2xy$,

$$y' = \frac{y^2 - x^2}{2xy} = \frac{y}{2x} - \frac{x}{2y}.$$

Now substitute y and y' from (9) and then simplify by subtracting u on both sides,

$$u'x + u = \frac{u}{2} - \frac{1}{2u}, \qquad u'x = -\frac{u}{2} - \frac{1}{2u} = \frac{-u^2 - 1}{2u}.$$

You see that in the last equation you can now separate the variables,

$$\frac{2u\,du}{1 + u^2} = -\frac{dx}{x}. \qquad \text{By integration,} \qquad \ln(1 + u^2) = -\ln|x| + c^* = \ln\left|\frac{1}{x}\right| + c^*.$$

Take exponents on both sides to get $1 + u^2 = c/x$ or $1 + (y/x)^2 = c/x$. Multiply the last equation by x^2 to obtain (Fig. 14)

$$x^2 + y^2 = cx. \qquad \text{Thus} \qquad \left(x - \frac{c}{2}\right)^2 + y^2 = \frac{c^2}{4}.$$

This general solution represents a family of circles passing through the origin with centers on the x-axis. ■

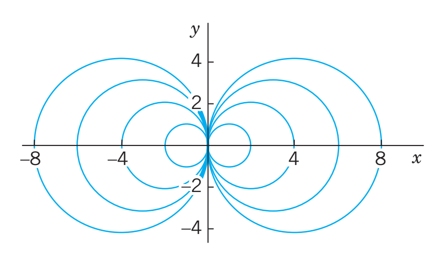


Fig. 14. General solution (family of circles) in Example 8

PROBLEM SET 1.3

1. **CAUTION! Constant of integration.** Why is it important to introduce the constant of integration immediately when you integrate?

2–10 GENERAL SOLUTION

Find a general solution. Show the steps of derivation. Check your answer by substitution.

2. $y^3y' + x^3 = 0$
3. $y' = \sec^2 y$
4. $y' \sin 2\pi x = \pi y \cos 2\pi x$
5. $yy' + 36x = 0$
6. $y' = e^{2x-1}y^2$
7. $xy' = y + 2x^3 \sin^2 \dfrac{y}{x}$ (Set $y/x = u$)
8. $y' = (y + 4x)^2$ (Set $y + 4x = v$)
9. $xy' = y^2 + y$ (Set $y/x = u$)
10. $xy' = x + y$ (Set $y/x = u$)

11–17 INITIAL VALUE PROBLEMS (IVPs)

Solve the IVP. Show the steps of derivation, beginning with the general solution.

11. $xy' + y = 0, \quad y(4) = 6$
12. $y' = 1 + 4y^2, \quad y(1) = 0$
13. $y'\cosh^2 x = \sin^2 y, \quad y(0) = \frac{1}{2}\pi$
14. $dr/dt = -2tr, \quad r(0) = r_0$
15. $y' = -4x/y, \quad y(2) = 3$
16. $y' = (x + y - 2)^2, \quad y(0) = 2$ (Set $v = x + y - 2$)
17. $xy' = y + 3x^4 \cos^2 (y/x), \quad y(1) = 0$ (Set $y/x = u$)
18. **Particular solution.** Introduce limits of integration in (3) such that y obtained from (3) satisfies the initial condition $y(x_0) = y_0$.

19–36 MODELING, APPLICATIONS

19. Exponential growth. If the growth rate of the number of bacteria at any time t is proportional to the number present at t and doubles in 1 week, how many bacteria can be expected after 2 weeks? After 4 weeks?

20. Another population model.

(a) If the birth rate and death rate of the number of bacteria are proportional to the number of bacteria present, what is the population as a function of time.

(b) What is the limiting situation for increasing time? Interpret it.

21. Radiocarbon dating. What should be the ${}^{14}_{6}C$ content (in percent of y_0) of a fossilized tree that is claimed to be 3000 years old? (See Example 4.)

22. Linear accelerators are used in physics for accelerating charged particles. Suppose that an alpha particle enters an accelerator and undergoes a constant acceleration that increases the speed of the particle from 10^3 m/sec to 10^4 m/sec in 10^{-3} sec. Find the acceleration a and the distance traveled during that period of 10^{-3} sec.

23. Boyle–Mariotte's law for ideal gases.[5] Experiments show for a gas at low pressure p (and constant temperature) the rate of change of the volume $V(p)$ equals $-V/p$. Solve the model.

24. Mixing problem. A tank contains 400 gal of brine in which 100 lb of salt are dissolved. Fresh water runs into the tank at a rate of 2 gal/min.The mixture, kept practically uniform by stirring, runs out at the same rate. How much salt will there be in the tank at the end of 1 hour?

25. Newton's law of cooling. A thermometer, reading 5°C, is brought into a room whose temperature is 22°C. One minute later the thermometer reading is 12°C. How long does it take until the reading is practically 22°C, say, 21.9°C?

26. Gompertz growth in tumors. The Gompertz model is $y' = -Ay \ln y$ $(A > 0)$, where $y(t)$ is the mass of tumor cells at time t. The model agrees well with clinical observations. The declining growth rate with increasing $y > 1$ corresponds to the fact that cells in the interior of a tumor may die because of insufficient oxygen and nutrients. Use the ODE to discuss the growth and decline of solutions (tumors) and to find constant solutions. Then solve the ODE.

27. Dryer. If a wet sheet in a dryer loses its moisture at a rate proportional to its moisture content, and if it loses half of its moisture during the first 10 min of drying, when will it be practically dry, say, when will it have lost 99% of its moisture? First guess, then calculate.

28. Estimation. Could you see, practically without calculation, that the answer in Prob. 27 must lie between 60 and 70 min? Explain.

29. Alibi? Jack, arrested when leaving a bar, claims that he has been inside for at least half an hour (which would provide him with an alibi). The police check the water temperature of his car (parked near the entrance of the bar) at the instant of arrest and again 30 min later, obtaining the values 190°F and 110°F, respectively. Do these results give Jack an alibi? (Solve by inspection.)

30. Rocket. A rocket is shot straight up from the earth, with a net acceleration (= acceleration by the rocket engine minus gravitational pullback) of $7t$ m/sec^2 during the initial stage of flight until the engine cut out at $t = 10$ sec. How high will it go, air resistance neglected?

31. Solution curves of $y' = g(y/x)$. Show that any (nonvertical) straight line through the origin of the xy-plane intersects all these curves of a given ODE at the same angle.

32. Friction. If a body slides on a surface, it experiences friction F (a force against the direction of motion). Experiments show that $|F| = \mu|N|$ (*Coulomb's*[6] *law of kinetic friction without lubrication*), where N is the normal force (force that holds the two surfaces together; see Fig. 15) and the constant of proportionality μ is called the *coefficient of kinetic friction*. In Fig. 15 assume that the body weighs 45 nt (about 10 lb; see front cover for conversion). $\mu = 0.20$ (corresponding to steel on steel), $a = 30°$, the slide is 10 m long, the initial velocity is zero, and air resistance is negligible. Find the velocity of the body at the end of the slide.

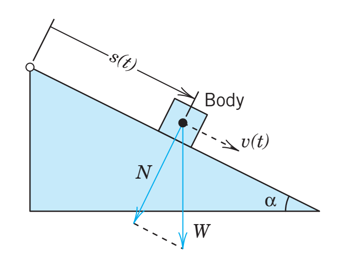


Fig. 15. Problem 32

[5]ROBERT BOYLE (1627–1691), English physicist and chemist, one of the founders of the Royal Society. EDME MARIOTTE (about 1620–1684), French physicist and prior of a monastry near Dijon. They found the law experimentally in 1662 and 1676, respectively.

[6]CHARLES AUGUSTIN DE COULOMB (1736–1806), French physicist and engineer.

33. Rope. To tie a boat in a harbor, how many times must a rope be wound around a bollard (a vertical rough cylindrical post fixed on the ground) so that a man holding one end of the rope can resist a force exerted by the boat 1000 times greater than the man can exert? First guess. Experiments show that the change ΔS of the force S in a small portion of the rope is proportional to S and to the small angle $\Delta\phi$ in Fig. 16. Take the proportionality constant 0.15. The result should surprise you!

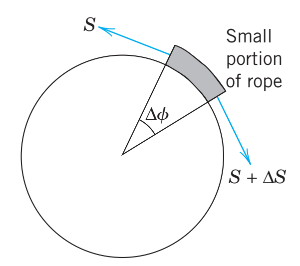


Fig. 16. Problem 33

34. TEAM PROJECT. Family of Curves. A family of curves can often be characterized as the general solution of $y' = f(x, y)$.

(a) Show that for the circles with center at the origin we get $y' = -x/y$.

(b) Graph some of the hyperbolas $xy = c$. Find an ODE for them.

(c) Find an ODE for the straight lines through the origin.

(d) You will see that the product of the right sides of the ODEs in (a) and (c) equals -1. Do you recognize this as the condition for the two families to be orthogonal (i.e., to intersect at right angles)? Do your graphs confirm this?

(e) Sketch families of curves of your own choice and find their ODEs. Can every family of curves be given by an ODE?

35. CAS PROJECT. Graphing Solutions. A CAS can usually graph solutions, even if they are integrals that cannot be evaluated by the usual analytical methods of calculus.

(a) Show this for the five initial value problems $y' = e^{-x^2}$, $y(0) = 0, \pm 1, \pm 2$ graphing all five curves on the same axes.

(b) Graph approximate solution curves, using the first few terms of the Maclaurin series (obtained by termwise integration of that of y') and compare with the exact curves.

(c) Repeat the work in (a) for another ODE and initial conditions of your own choice, leading to an integral that cannot be evaluated as indicated.

36. TEAM PROJECT. Torricelli's Law. Suppose that the tank in Example 7 is hemispherical, of radius R, initially full of water, and has an outlet of 5 cm^2 cross-sectional area at the bottom. (Make a sketch.) Set up the model for outflow. Indicate what portion of your work in Example 7 you can use (so that it can become part of the general method independent of the shape of the tank). Find the time t to empty the tank (a) for any R, (b) for $R = 1$ m. Plot t as function of R. Find the time when $h = R/2$ (a) for any R, (b) for $R = 1$ m.

1.4 Exact ODEs. Integrating Factors

We recall from calculus that if a function $u(x, y)$ has continuous partial derivatives, its **differential** (also called its *total differential*) is

$$du = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy.$$

From this it follows that if $u(x, y) = c = \text{const}$, then $du = 0$.

For example, if $u = x + x^2y^3 = c$, then

$$du = (1 + 2xy^3)\,dx + 3x^2y^2\,dy = 0$$

or

$$y' = \frac{dy}{dx} = -\frac{1 + 2xy^3}{3x^2y^2},$$

an ODE that we can solve by going backward. This idea leads to a powerful solution method as follows.

A first-order ODE $M(x, y) + N(x, y)y' = 0$, written as (use $dy = y' dx$ as in Sec. 1.3)

$$M(x, y)\,dx + N(x, y)\,dy = 0 \tag{1}$$

is called an **exact differential equation** if the **differential form** $M(x, y)\,dx + N(x, y)\,dy$ is **exact**, that is, this form is the differential

$$du = \frac{\partial u}{\partial x}dx + \frac{\partial u}{\partial y}dy \tag{2}$$

of some function $u(x, y)$. Then (1) can be written

$$du = 0.$$

By integration we immediately obtain the general solution of (1) in the form

$$u(x, y) = c. \tag{3}$$

This is called an **implicit solution**, in contrast to a solution $y = h(x)$ as defined in Sec. 1.1, which is also called an *explicit solution,* for distinction. Sometimes an implicit solution can be converted to explicit form. (Do this for $x^2 + y^2 = 1$.) If this is not possible, your CAS may graph a figure of the **contour lines** (3) of the function $u(x, y)$ and help you in understanding the solution.

Comparing (1) and (2), we see that (1) is an exact differential equation if there is some function $u(x, y)$ such that

$$\text{(a)}\quad \frac{\partial u}{\partial x} = M, \qquad \text{(b)}\quad \frac{\partial u}{\partial y} = N. \tag{4}$$

From this we can derive a formula for checking whether (1) is exact or not, as follows.

Let M and N be continuous and have continuous first partial derivatives in a region in the xy-plane whose boundary is a closed curve without self-intersections. Then by partial differentiation of (4) (see App. 3.2 for notation),

$$\frac{\partial M}{\partial y} = \frac{\partial^2 u}{\partial y\,\partial x},$$

$$\frac{\partial N}{\partial x} = \frac{\partial^2 u}{\partial x\,\partial y}.$$

By the assumption of continuity the two second partial derivaties are equal. Thus

$$\frac{\partial M}{\partial y} = \frac{\partial N}{\partial x}. \tag{5}$$

This condition is not only necessary but also sufficient for (1) to be an exact differential equation. (We shall prove this in Sec. 10.2 in another context. Some calculus books, for instance, [GenRef 12], also contain a proof.)

If (1) is exact, the function $u(x, y)$ can be found by inspection or in the following systematic way. From (4a) we have by integration with respect to x

(6) $$u = \int M\,dx + k(y);$$

in this integration, y is to be regarded as a constant, and $k(y)$ plays the role of a "constant" of integration. To determine $k(y)$, we derive $\partial u/\partial y$ from (6), use (4b) to get dk/dy, and integrate dk/dy to get k. (See Example 1, below.)

Formula (6) was obtained from (4a). Instead of (4a) we may equally well use (4b). Then, instead of (6), we first have by integration with respect to y

(6*) $$u = \int N\,dy + l(x).$$

To determine $l(x)$, we derive $\partial u/\partial x$ from (6*), use (4a) to get dl/dx, and integrate. We illustrate all this by the following typical examples.

EXAMPLE 1 **An Exact ODE**

Solve

(7) $$\cos(x + y)\,dx + (3y^2 + 2y + \cos(x + y))\,dy = 0.$$

Solution. ***Step 1. Test for exactness.*** Our equation is of the form (1) with

$$M = \cos(x + y),$$

$$N = 3y^2 + 2y + \cos(x + y).$$

Thus

$$\frac{\partial M}{\partial y} = -\sin(x + y),$$

$$\frac{\partial N}{\partial x} = -\sin(x + y).$$

From this and (5) we see that (7) is exact.

Step 2. Implicit general solution. From (6) we obtain by integration

(8) $$u = \int M\,dx + k(y) = \int \cos(x + y)\,dx + k(y) = \sin(x + y) + k(y).$$

To find $k(y)$, we differentiate this formula with respect to y and use formula (4b), obtaining

$$\frac{\partial u}{\partial y} = \cos(x + y) + \frac{dk}{dy} = N = 3y^2 + 2y + \cos(x + y).$$

Hence $dk/dy = 3y^2 + 2y$. By integration, $k = y^3 + y^2 + c^*$. Inserting this result into (8) and observing (3), we obtain the *answer*

$$u(x, y) = \sin(x + y) + y^3 + y^2 = c.$$

Step 3. Checking an implicit solution. We can check by differentiating the implicit solution $u(x, y) = c$ implicitly and see whether this leads to the given ODE (7):

$$du = \frac{\partial u}{\partial x}\,dx + \frac{\partial u}{\partial y}\,dy = \cos(x + y)\,dx + (\cos(x + y) + 3y^2 + 2y)\,dy = 0. \tag{9}$$

This completes the check. ■

EXAMPLE 2 An Initial Value Problem

Solve the initial value problem

$$(\cos y \sinh x + 1)\,dx - \sin y \cosh x\,dy = 0, \qquad y(1) = 2. \tag{10}$$

Solution. You may verify that the given ODE is exact. We find u. For a change, let us use (6*),

$$u = -\int \sin y \cosh x\,dy + l(x) = \cos y \cosh x + l(x).$$

From this, $\partial u/\partial x = \cos y \sinh x + dl/dx = M = \cos y \sinh x + 1$. Hence $dl/dx = 1$. By integration, $l(x) = x + c^*$. This gives the general solution $u(x, y) = \cos y \cosh x + x = c$. From the initial condition, $\cos 2 \cosh 1 + 1 = 0.358 = c$. Hence the answer is $\cos y \cosh x + x = 0.358$. Figure 17 shows the particular solutions for $c = 0, 0.358$ (thicker curve), 1, 2, 3. Check that the answer satisfies the ODE. (Proceed as in Example 1.) Also check that the initial condition is satisfied. ■

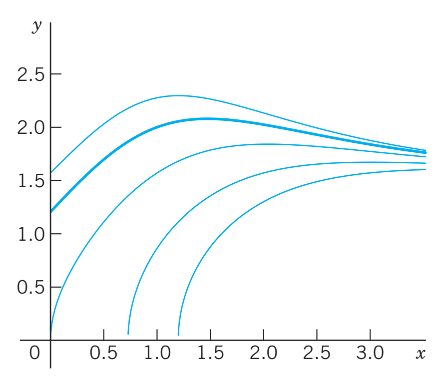


Fig. 17. Particular solutions in Example 2

EXAMPLE 3 WARNING! Breakdown in the Case of Nonexactness

The equation $-y\,dx + x\,dy = 0$ is not exact because $M = -y$ and $N = x$, so that in (5), $\partial M/\partial y = -1$ but $\partial N/\partial x = 1$. Let us show that in such a case the present method does not work. From (6),

$$u = \int M\,dx + k(y) = -xy + k(y), \qquad \text{hence} \qquad \frac{\partial u}{\partial y} = -x + \frac{dk}{dy}.$$

Now, $\partial u/\partial y$ should equal $N = x$, by (4b). However, this is impossible because $k(y)$ can depend only on y. Try (6*); it will also fail. Solve the equation by another method that we have discussed. ■

Reduction to Exact Form. Integrating Factors

The ODE in Example 3 is $-y\,dx + x\,dy = 0$. It is not exact. However, if we multiply it by $1/x^2$, we get an exact equation [check exactness by (5)!],

$$\frac{-y\,dx + x\,dy}{x^2} = -\frac{y}{x^2}\,dx + \frac{1}{x}\,dy = d\left(\frac{y}{x}\right) = 0. \tag{11}$$

Integration of (11) then gives the general solution $y/x = c = \text{const}$.

This example gives the idea. All we did was to multiply a given nonexact equation, say,

$$P(x, y)\,dx + Q(x, y)\,dy = 0, \tag{12}$$

by a function F that, in general, will be a function of both x and y. The result was an equation

$$FP\,dx + FQ\,dy = 0 \tag{13}$$

that is exact, so we can solve it as just discussed. Such a function $F(x, y)$ is then called an **integrating factor** of (12).

EXAMPLE 4 Integrating Factor

The integrating factor in (11) is $F = 1/x^2$. Hence in this case the exact equation (13) is

$$FP\,dx + FQ\,dy = \frac{-y\,dx + x\,dy}{x^2} = d\left(\frac{y}{x}\right) = 0. \qquad \text{Solution} \qquad \frac{y}{x} = c.$$

These are straight lines $y = cx$ through the origin. (Note that $x = 0$ is also a solution of $-y\,dx + x\,dy = 0$.)

It is remarkable that we can readily find other integrating factors for the equation $-y\,dx + x\,dy = 0$, namely, $1/y^2$, $1/(xy)$, and $1/(x^2 + y^2)$, because

$$\frac{-y\,dx + x\,dy}{y^2} = d\left(\frac{x}{y}\right), \quad \frac{-y\,dx + x\,dy}{xy} = -d\left(\ln\frac{x}{y}\right), \quad \frac{-y\,dx + x\,dy}{x^2 + y^2} = d\left(\arctan\frac{y}{x}\right). \tag{14}$$ ■

How to Find Integrating Factors

In simpler cases we may find integrating factors by inspection or perhaps after some trials, keeping (14) in mind. In the general case, the idea is the following.

For $M\,dx + N\,dy = 0$ the exactness condition (5) is $\partial M/\partial y = \partial N/\partial x$. Hence for (13), $FP\,dx + FQ\,dy = 0$, the exactness condition is

$$\frac{\partial}{\partial y}(FP) = \frac{\partial}{\partial x}(FQ). \tag{15}$$

By the product rule, with subscripts denoting partial derivatives, this gives

$$F_yP + FP_y = F_xQ + FQ_x.$$

In the general case, this would be complicated and useless. So we follow the ***Golden Rule:*** *If you cannot solve your problem, try to solve a simpler one*—the result may be useful (and may also help you later on). Hence we look for an integrating factor depending only on ***one*** variable: fortunately, in many practical cases, there are such factors, as we shall see. Thus, let $F = F(x)$. Then $F_y = 0$, and $F_x = F' = dF/dx$, so that (15) becomes

$$FP_y = F'Q + FQ_x.$$

Dividing by FQ and reshuffling terms, we have

$$\frac{1}{F}\frac{dF}{dx} = R, \qquad \text{where} \qquad R = \frac{1}{Q}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right). \tag{16}$$

This proves the following theorem.

THEOREM 1

Integrating Factor $F(x)$

If (12) is such that the right side R of (16) depends only on x, then (12) has an integrating factor $F = F(x)$, which is obtained by integrating (16) and taking exponents on both sides.

$$F(x) = \exp \int R(x)\, dx. \qquad (17)$$

Similarly, if $F^* = F^*(y)$, then instead of (16) we get

$$\frac{1}{F^*}\frac{dF^*}{dy} = R^*, \qquad \text{where} \qquad R^* = \frac{1}{P}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) \qquad (18)$$

and we have the companion

THEOREM 2

Integrating Factor $F^*(y)$

If (12) is such that the right side R^* of (18) depends only on y, then (12) has an integrating factor $F^* = F^*(y)$, which is obtained from (18) in the form

$$F^*(y) = \exp \int R^*(y)\, dy. \qquad (19)$$

EXAMPLE 5

Application of Theorems 1 and 2. Initial Value Problem

Using Theorem 1 or 2, find an integrating factor and solve the initial value problem

$$(e^{x+y} + ye^y)\, dx + (xe^y - 1)\, dy = 0, \quad y(0) = -1 \qquad (20)$$

Solution. ***Step 1. Nonexactness.*** The exactness check fails:

$$\frac{\partial P}{\partial y} = \frac{\partial}{\partial y}(e^{x+y} + ye^y) = e^{x+y} + e^y + ye^y \qquad \text{but} \qquad \frac{\partial Q}{\partial x} = \frac{\partial}{\partial x}(xe^y - 1) = e^y.$$

Step 2. Integrating factor. General solution. Theorem 1 fails because R [the right side of (16)] depends on both x and y.

$$R = \frac{1}{Q}\left(\frac{\partial P}{\partial y} - \frac{\partial Q}{\partial x}\right) = \frac{1}{xe^y - 1}(e^{x+y} + e^y + ye^y - e^y).$$

Try Theorem 2. The right side of (18) is

$$R^* = \frac{1}{P}\left(\frac{\partial Q}{\partial x} - \frac{\partial P}{\partial y}\right) = \frac{1}{e^{x+y} + ye^y}(e^y - e^{x+y} - e^y - ye^y) = -1.$$

Hence (19) gives the integrating factor $F^*(y) = e^{-y}$. From this result and (20) you get the exact equation

$$(e^x + y)\, dx + (x - e^{-y})\, dy = 0.$$

Test for exactness; you will get 1 on both sides of the exactness condition. By integration, using (4a),

$$u = \int (e^x + y)\,dx = e^x + xy + k(y).$$

Differentiate this with respect to y and use (4b) to get

$$\frac{\partial u}{\partial y} = x + \frac{dk}{dy} = N = x - e^{-y}, \qquad \frac{dk}{dy} = -e^{-y}, \qquad k = e^{-y} + c^*.$$

Hence the general solution is

$$u(x, y) = e^x + xy + e^{-y} = c.$$

Setp 3. Particular solution. The initial condition $y(0) = -1$ gives $u(0, -1) = 1 + 0 + e = 3.72$. Hence the answer is $e^x + xy + e^{-y} = 1 + e = 3.72$. Figure 18 shows several particular solutions obtained as level curves of $u(x, y) = c$, obtained by a CAS, a convenient way in cases in which it is impossible or difficult to cast a solution into explicit form. Note the curve that (nearly) satisfies the initial condition.

Step 4. Checking. Check by substitution that the answer satisfies the given equation as well as the initial condition. ■

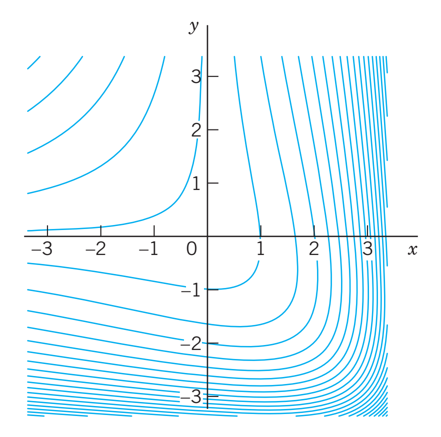


Fig. 18. Particular solutions in Example 5

PROBLEM SET 1.4

1–14 ODEs. INTEGRATING FACTORS

Test for exactness. If exact, solve. If not, use an integrating factor as given or obtained by inspection or by the theorems in the text. Also, if an initial condition is given, find the corresponding particular solution.

1. $2xy\,dx + x^2\,dy = 0$
2. $x^3 dx + y^3 dy = 0$
3. $\sin x \cos y\,dx + \cos x \sin y\,dy = 0$
4. $e^{3\theta}(dr + 3r\,d\theta) = 0$
5. $(x^2 + y^2)\,dx - 2xy\,dy = 0$
6. $3(y + 1)\,dx = 2x\,dy, \quad (y + 1)x^{-4}$
7. $2x \tan y\,dx + \sec^2 y\,dy = 0$
8. $e^x(\cos y\,dx - \sin y\,dy) = 0$
9. $e^{2x}(2\cos y\,dx - \sin y\,dy) = 0, \quad y(0) = 0$
10. $y\,dx + [y + \tan(x + y)]\,dy = 0, \quad \cos(x + y)$
11. $2\cosh x \cos y\,dx = \sinh x \sin y\,dy$
12. $(2xy\,dx + dy)e^{x^2} = 0, \quad y(0) = 2$
13. $e^{-y}\,dx + e^{-x}(-e^{-y} + 1)\,dy = 0, \quad F = e^{x+y}$
14. $(a + 1)y\,dx + (b + 1)x\,dy = 0, \quad y(1) = 1,$ $F = x^a y^b$
15. **Exactness.** Under what conditions for the constants a, b, k, l is $(ax + by)\,dx + (kx + ly)\,dy = 0$ exact? Solve the exact ODE.

16. TEAM PROJECT. Solution by Several Methods. Show this as indicated. Compare the amount of work.

(a) $e^y(\sinh x\,dx + \cosh x\,dy) = 0$ as an exact ODE and by separation.

(b) $(1 + 2x)\cos y\,dx + dy/\cos y = 0$ by Theorem 2 and by separation.

(c) $(x^2 + y^2)\,dx - 2xy\,dy = 0$ by Theorem 1 or 2 and by separation with $v = y/x$.

(d) $3x^2 y\,dx + 4x^3\,dy = 0$ by Theorems 1 and 2 and by separation.

(e) Search the text and the problems for further ODEs that can be solved by more than one of the methods discussed so far. Make a list of these ODEs. Find further cases of your own.

17. WRITING PROJECT. Working Backward. Working backward from the solution to the problem is useful in many areas. Euler, Lagrange, and other great masters did it. To get additional insight into the idea of integrating factors, start from a $u(x, y)$ of your choice, find $du = 0$, destroy exactness by division by some $F(x, y)$, and see what ODE's solvable by integrating factors you can get. Can you proceed systematically, beginning with the simplest $F(x, y)$?

18. CAS PROJECT. Graphing Particular Solutions. Graph particular solutions of the following ODE, proceeding as explained.

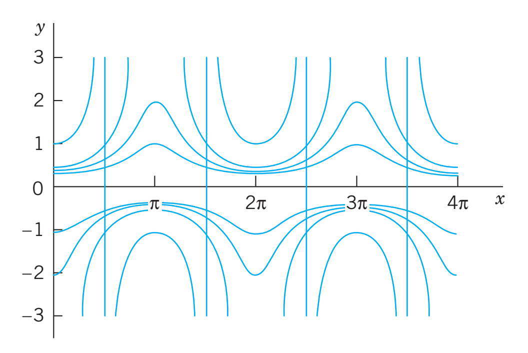
(21) $$dy - y^2 \sin x\,dx = 0.$$

(a) Show that (21) is not exact. Find an integrating factor using either Theorem 1 or 2. Solve (21).

(b) Solve (21) by separating variables. Is this simpler than (a)?

(c) Graph the seven particular solutions satisfying the following initial conditions $y(0) = 1$, $y(\pi/2) = \pm\frac{1}{2}$, $\pm\frac{2}{3}$, ± 1 (see figure below).

(d) Which solution of (21) do we not get in (a) or (b)?



Particular solutions in CAS Project 18

1.5 Linear ODEs. Bernoulli Equation. Population Dynamics

Linear ODEs or ODEs that can be transformed to linear form are models of various phenomena, for instance, in physics, biology, population dynamics, and ecology, as we shall see. A first-order ODE is said to be **linear** if it can be brought into the form

(1) $$y' + p(x)y = r(x),$$

by algebra, and **nonlinear** if it cannot be brought into this form.

The defining feature of the linear ODE (1) is that it is linear in both the unknown function y and its derivative $y' = dy/dx$, whereas p and r may be ***any*** given functions of x. If in an application the independent variable is time, we write t instead of x.

If the first term is $f(x)y'$ (instead of y'), divide the equation by $f(x)$ to get the **standard form** (1), with y' as the first term, which is practical.

For instance, $y' \cos x + y \sin x = x$ is a linear ODE, and its standard form is $y' + y \tan x = x \sec x$.

The function $r(x)$ on the right may be a force, and the solution $y(x)$ a displacement in a motion or an electrical current or some other physical quantity. In engineering, $r(x)$ is frequently called the **input**, and $y(x)$ is called the **output** or the *response* to the input (and, if given, to the initial condition).

Homogeneous Linear ODE. We want to solve (1) in some interval $a < x < b$, call it J, and we begin with the simpler special case that $r(x)$ is zero for all x in J. (This is sometimes written $r(x) \equiv 0$.) Then the ODE (1) becomes

$$y' + p(x)y = 0 \qquad \textbf{(2)}$$

and is called **homogeneous**. By separating variables and integrating we then obtain

$$\frac{dy}{y} = -p(x)\,dx, \qquad \text{thus} \qquad \ln|y| = -\int p(x)\,dx + c^*.$$

Taking exponents on both sides, we obtain the general solution of the homogeneous ODE (2),

$$y(x) = ce^{-\int p(x)dx} \qquad (c = \pm e^{c^*} \quad \text{when} \quad y \gtrless 0); \qquad \textbf{(3)}$$

here we may also choose $c = 0$ and obtain the **trivial solution** $y(x) = 0$ for all x in that interval.

Nonhomogeneous Linear ODE. We now solve (1) in the case that $r(x)$ in (1) is not everywhere zero in the interval J considered. Then the ODE (1) is called **nonhomogeneous**. It turns out that in this case, (1) has a pleasant property; namely, it has an integrating factor depending only on x. We can find this factor $F(x)$ by Theorem 1 in the previous section or we can proceed directly, as follows. We multiply (1) by $F(x)$, obtaining

$$Fy' + pFy = rF. \qquad (1^*)$$

The left side is the derivative $(Fy)' = F'y + Fy'$ of the product Fy if

$$pFy = F'y, \qquad \text{thus} \qquad pF = F'.$$

By separating variables, $dF/F = p\,dx$. By integration, writing $h = \int p\,dx$,

$$\ln|F| = h = \int p\,dx, \qquad \text{thus} \qquad F = e^h.$$

With this F and $h' = p$, Eq. (1*) becomes

$$e^h y' + h'e^h y = e^h y' + (e^h)'y = (e^h y)' = re^h.$$

By integration,

$$e^h y = \int e^h r\,dx + c.$$

Dividing by e^h, we obtain the desired solution formula

$$y(x) = e^{-h}\left(\int e^h r\,dx + c\right), \qquad h = \int p(x)\,dx. \qquad \textbf{(4)}$$

This reduces solving (1) to the generally simpler task of evaluating integrals. For ODEs for which this is still difficult, you may have to use a numeric method for integrals from Sec. 19.5 or for the ODE itself from Sec. 21.1. We mention that h has nothing to do with $h(x)$ in Sec. 1.1 and that the constant of integration in h does not matter; see Prob. 2.

The structure of (4) is interesting. The only quantity depending on a given initial condition is c. Accordingly, writing (4) as a sum of two terms,

$$y(x) = e^{-h} \int e^{h} r \, dx + c e^{-h}, \qquad (4^*)$$

we see the following:

$$\text{Total Output} = \text{Response to the Input } r + \text{Response to the Initial Data.} \qquad \textbf{(5)}$$

EXAMPLE 1 **First-Order ODE, General Solution, Initial Value Problem**

Solve the initial value problem

$$y' + y \tan x = \sin 2x, \qquad y(0) = 1.$$

Solution. Here $p = \tan x$, $r = \sin 2x = 2 \sin x \cos x$, and

$$h = \int p \, dx = \int \tan x \, dx = \ln |\sec x|.$$

From this we see that in (4),

$$e^{h} = \sec x, \qquad e^{-h} = \cos x, \qquad e^{h} r = (\sec x)(2 \sin x \cos x) = 2 \sin x,$$

and the general solution of our equation is

$$y(x) = \cos x \left(2 \int \sin x \, dx + c \right) = c \cos x - 2 \cos^2 x.$$

From this and the initial condition, $1 = c \cdot 1 - 2 \cdot 1^2$; thus $c = 3$ and the solution of our initial value problem is $y = 3 \cos x - 2 \cos^2 x$. Here $3 \cos x$ is the response to the initial data, and $-2 \cos^2 x$ is the response to the input $\sin 2x$. ■

EXAMPLE 2 **Electric Circuit**

Model the ***RL*-circuit** in Fig. 19 and solve the resulting ODE for the current $I(t)$ A (amperes), where t is time. Assume that the circuit contains as an **EMF** $E(t)$ (electromotive force) a battery of $E = 48$ V (volts), which is constant, a ***resistor*** of $R = 11\ \Omega$ (ohms), and an ***inductor*** of $L = 0.1$ H (henrys), and that the current is initially zero.

Physical Laws. A current I in the circuit causes a **voltage drop** RI across the resistor **(Ohm's law)** and a voltage drop $LI' = L\, dI/dt$ across the conductor, and the sum of these two voltage drops equals the EMF **(Kirchhoff's Voltage Law, KVL)**.

Remark. In general, KVL states that "The voltage (the electromotive force EMF) impressed on a closed loop is equal to the sum of the voltage drops across all the other elements of the loop." For Kirchoff's Current Law (KCL) and historical information, see footnote 7 in Sec. 2.9.

Solution. According to these laws the model of the RL-circuit is $LI' + RI = E(t)$, in standard form

$$I' + \frac{R}{L} I = \frac{E(t)}{L}. \qquad (6)$$

We can solve this linear ODE by (4) with $x = t$, $y = I$, $p = R/L$, $h = (R/L)t$, obtaining the general solution

$$I = e^{-(R/L)t}\left(\int e^{(R/L)t}\frac{E(t)}{L}\,dt + c\right).$$

By integration,

$$I = e^{-(R/L)t}\left(\frac{E}{L}\,\frac{e^{(R/L)t}}{R/L} + c\right) = \frac{E}{R} + ce^{-(R/L)t}. \tag{7}$$

In our case, $R/L = 11/0.1 = 110$ and $E(t) = 48/0.1 = 480 = \text{const}$; thus,

$$I = \tfrac{48}{11} + ce^{-110t}.$$

In modeling, one often gets better insight into the nature of a solution (and smaller roundoff errors) by inserting given numeric data only near the end. Here, the general solution (7) shows that the current approaches the limit $E/R = 48/11$ faster the larger R/L is, in our case, $R/L = 11/0.1 = 110$, and the approach is very fast, from below if $I(0) < 48/11$ or from above if $I(0) > 48/11$. If $I(0) = 48/11$, the solution is constant (48/11 A). See Fig. 19.

The initial value $I(0) = 0$ gives $I(0) = E/R + c = 0$, $c = -E/R$ and the particular solution

$$I = \frac{E}{R}(1 - e^{-(R/L)t}), \qquad \text{thus} \qquad I = \frac{48}{11}(1 - e^{-110t}). \tag{8}$$

■

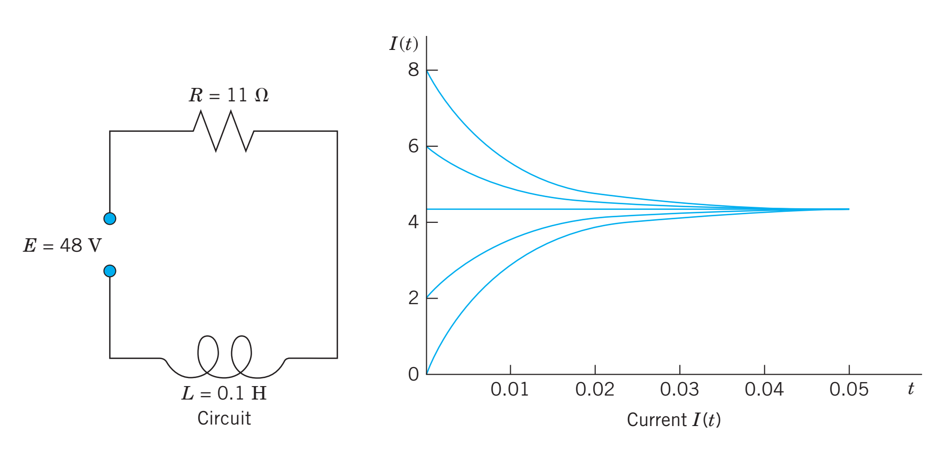


Fig. 19. *RL*-circuit

EXAMPLE 3 Hormone Level

Assume that the level of a certain hormone in the blood of a patient varies with time. Suppose that the time rate of change is the difference between a sinusoidal input of a 24-hour period from the thyroid gland and a continuous removal rate proportional to the level present. Set up a model for the hormone level in the blood and find its general solution. Find the particular solution satisfying a suitable initial condition.

Solution. ***Step 1. Setting up a model.*** Let $y(t)$ be the hormone level at time t. Then the removal rate is $Ky(t)$. The input rate is $A + B\cos\omega t$, where $\omega = 2\pi/24 = \pi/12$ and A is the average input rate; here $A \geqq B$ to make the input rate nonnegative. The constants A, B, K can be determined from measurements. Hence the model is the linear ODE

$$y'(t) = \text{In} - \text{Out} = A + B\cos\omega t - Ky(t), \qquad \text{thus} \qquad y' + Ky = A + B\cos\omega t.$$

The initial condition for a particular solution y_{part} is $y_{\text{part}}(0) = y_0$ with $t = 0$ suitably chosen, for example, 6:00 A.M.

Step 2. General solution. In (4) we have $p = K = \text{const}$, $h = Kt$, and $r = A + B\cos\omega t$. Hence (4) gives the general solution (evaluate $\int e^{Kt}\cos\omega t\,dt$ by integration by parts)

$$y(t) = e^{-Kt}\int e^{Kt}\left(A + B\cos\omega t\right)dt + ce^{-Kt}$$

$$= e^{-Kt}e^{Kt}\left[\frac{A}{K} + \frac{B}{K^2 + \omega^2}\left(K\cos\omega t + \omega\sin\omega t\right)\right] + ce^{-Kt}$$

$$= \frac{A}{K} + \frac{B}{K^2 + (\pi/12)^2}\left(K\cos\frac{\pi t}{12} + \frac{\pi}{12}\sin\frac{\pi t}{12}\right) + ce^{-Kt}.$$

The last term decreases to 0 as t increases, practically after a short time and regardless of c (that is, of the initial condition). The other part of $y(t)$ is called the **steady-state solution** because it consists of constant and periodic terms. The entire solution is called the **transient-state solution** because it models the transition from rest to the steady state. These terms are used quite generally for physical and other systems whose behavior depends on time.

Step 3. Particular solution. Setting $t = 0$ in $y(t)$ and choosing $y_0 = 0$, we have

$$y(0) = \frac{A}{K} + \frac{B}{K^2 + (\pi/12)^2}\frac{u}{\pi}K + c = 0, \qquad \text{thus} \qquad c = -\frac{A}{K} - \frac{KB}{K^2 + (\pi/12)^2}.$$

Inserting this result into $y(t)$, we obtain the particular solution

$$y_{\text{part}}(t) = \frac{A}{K} + \frac{B}{K^2 + (\pi/12)^2}\left(K\cos\frac{\pi t}{12} + \frac{\pi}{12}\sin\frac{\pi t}{12}\right) - \left(\frac{A}{K} + \frac{KB}{K^2 + (\pi/12)^2}\right)e^{-K}$$

with the steady-state part as before. To plot y_{part} we must specify values for the constants, say, $A = B = 1$ and $K = 0.05$. Figure 20 shows this solution. Notice that the transition period is relatively short (although K is small), and the curve soon looks sinusoidal; this is the response to the input $A + B\cos\left(\frac{1}{12}\pi t\right) = 1 + \cos\left(\frac{1}{12}\pi t\right)$. ■

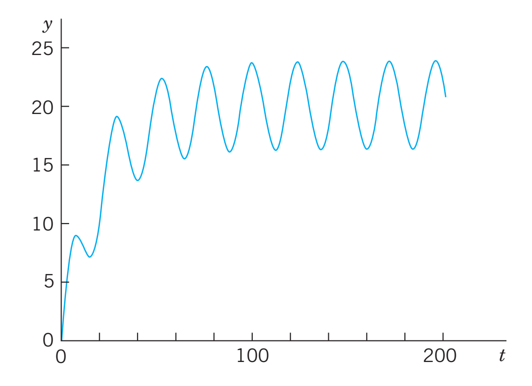


Fig. 20. Particular solution in Example 3

Reduction to Linear Form. Bernoulli Equation

Numerous applications can be modeled by ODEs that are nonlinear but can be transformed to linear ODEs. One of the most useful ones of these is the **Bernoulli equation**[7]

(9) $$y' + p(x)y = g(x)y^a \qquad (a \text{ any real number}).$$

[7]JAKOB BERNOULLI (1654–1705), Swiss mathematician, professor at Basel, also known for his contribution to elasticity theory and mathematical probability. The method for solving Bernoulli's equation was discovered by Leibniz in 1696. Jakob Bernoulli's students included his nephew NIKLAUS BERNOULLI (1687–1759), who contributed to probability theory and infinite series, and his youngest brother JOHANN BERNOULLI (1667–1748), who had profound influence on the development of calculus, became Jakob's successor at Basel, and had among his students GABRIEL CRAMER (see Sec. 7.7) and LEONHARD EULER (see Sec. 2.5). His son DANIEL BERNOULLI (1700–1782) is known for his basic work in fluid flow and the kinetic theory of gases.

If $a = 0$ or $a = 1$, Equation (9) is linear. Otherwise it is nonlinear. Then we set

$$u(x) = [y(x)]^{1-a}.$$

We differentiate this and substitute y' from (9), obtaining

$$u' = (1 - a)y^{-a}y' = (1 - a)y^{-a}(gy^{a} - py).$$

Simplification gives

$$u' = (1 - a)(g - py^{1-a}),$$

where $y^{1-a} = u$ on the right, so that we get the linear ODE

$$u' + (1 - a)pu = (1 - a)g. \tag{10}$$

For further ODEs reducible to linear form, see lnce's classic [A11] listed in App. 1. See also Team Project 30 in Problem Set 1.5.

EXAMPLE 4 Logistic Equation

Solve the following Bernoulli equation, known as the **logistic equation** (or **Verhulst equation**[8]):

$$y' = Ay - By^2 \tag{11}$$

Solution. Write (11) in the form (9), that is,

$$y' - Ay = -By^2$$

to see that $a = 2$, so that $u = y^{1-a} = y^{-1}$. Differentiate this u and substitute y' from (11),

$$u' = -y^{-2}y' = -y^{-2}(Ay - By^2) = B - Ay^{-1}.$$

The last term is $-Ay^{-1} = -Au$. Hence we have obtained the linear ODE

$$u' + Au = B.$$

The general solution is [by (4)]

$$u = ce^{-At} + B/A.$$

Since $u = 1/y$, this gives the general solution of (11),

$$y = \frac{1}{u} = \frac{1}{ce^{-At} + B/A} \tag{12}$$ (Fig. 21)

Directly from (11) we see that $y \equiv 0$ ($y(t) = 0$ for all t) is also a solution. ■

---

[8] PIERRE-FRANÇOIS VERHULST, Belgian statistician, who introduced Eq. (8) as a model for human population growth in 1838.

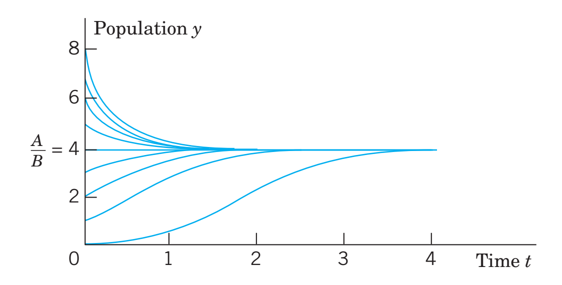


Fig. 21. Logistic population model. Curves (9) in Example 4 with $A/B = 4$

Population Dynamics

The logistic equation (11) plays an important role in **population dynamics**, a field that models the evolution of populations of plants, animals, or humans over time t. If $B = 0$, then (11) is $y' = dy/dt = Ay$. In this case its solution (12) is $y = (1/c)e^{At}$ and gives exponential growth, as for a small population in a large country (the United States in early times!). This is called ***Malthus's law***. (See also Example 3 in Sec. 1.1.)

The term $-By^2$ in (11) is a "braking term" that prevents the population from growing without bound. Indeed, if we write $y' = Ay[1 - (B/A)y]$, we see that if $y < A/B$, then $y' > 0$, so that an initially small population keeps growing as long as $y < A/B$. But if $y > A/B$, then $y' < 0$ and the population is decreasing as long as $y > A/B$. The limit is the same in both cases, namely, A/B. See Fig. 21.

We see that in the logistic equation (11) the independent variable t does not occur explicitly. An ODE $y' = f(t, y)$ in which t does not occur explicitly is of the form

$$y' = f(y) \qquad \textbf{(13)}$$

and is called an **autonomous ODE**. Thus the logistic equation (11) is autonomous.

Equation (13) has constant solutions, called **equilibrium solutions** or **equilibrium points**. These are determined by the zeros of $f(y)$, because $f(y) = 0$ gives $y' = 0$ by (13); hence $y = \text{const}$. These zeros are known as **critical points** of (13). An equilibrium solution is called **stable** if solutions close to it for some t remain close to it for all further t. It is called **unstable** if solutions initially close to it do not remain close to it as t increases. For instance, $y = 0$ in Fig. 21 is an unstable equilibrium solution, and $y = 4$ is a stable one. Note that (11) has the critical points $y = 0$ and $y = A/B$.

EXAMPLE 5 **Stable and Unstable Equilibrium Solutions. "Phase Line Plot"**

The ODE $y' = (y - 1)(y - 2)$ has the stable equilibrium solution $y_1 = 1$ and the unstable $y_2 = 2$, as the direction field in Fig. 22 suggests. The values y_1 and y_2 are the zeros of the parabola $f(y) = (y - 1)(y - 2)$ in the figure. Now, since the ODE is autonomous, we can "condense" the direction field to a "phase line plot" giving y_1 and y_2, and the direction (upward or downward) of the arrows in the field, and thus giving information about the stability or instability of the equilibrium solutions. ■

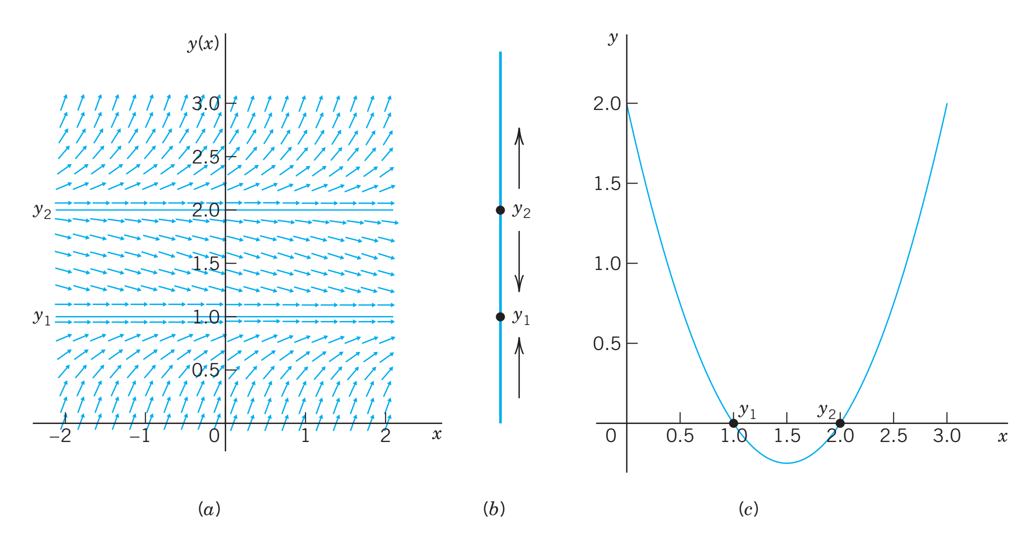


Fig. 22. Example 5. (A) Direction field. (B) "Phase line". (C) Parabola $f(y)$

A few further population models will be discussed in the problem set. For some more details of population dynamics, see C. W. Clark. *Mathematical Bioeconomics*: *The Mathematics of Conservation* 3rd ed. Hoboken, NJ, Wiley, 2010.

Further applications of linear ODEs follow in the next section.

PROBLEM SET 1.5

1. CAUTION! Show that $e^{-\ln x} = 1/x$ (not $-x$) and $e^{-\ln(\sec x)} = \cos x$.

2. Integration constant. Give a reason why in (4) you may choose the constant of integration in $\int p\, dx$ to be zero.

3–13 GENERAL SOLUTION. INITIAL VALUE PROBLEMS

Find the general solution. If an initial condition is given, find also the corresponding particular solution and graph or sketch it. (Show the details of your work.)

3. $y' - y = 5.2$

4. $y' = 2y - 4x$

5. $y' + ky = e^{-kx}$

6. $y' + 2y = 4 \cos 2x, \quad y(\frac{1}{4}\pi) = 3$

7. $xy' = 2y + x^3 e^x$

8. $y' + y \tan x = e^{-0.01x} \cos x, \quad y(0) = 0$

9. $y' + y \sin x = e^{\cos x}, \quad y(0) = -2.5$

10. $y' \cos x + (3y - 1) \sec x = 0, \quad y(\frac{1}{4}\pi) = 4/3$

11. $y' = (y - 2) \cot x$

12. $xy' + 4y = 8x^4, \quad y(1) = 2$

13. $y' = 6(y - 2.5) \tanh 1.5x$

14. CAS EXPERIMENT. (a) Solve the ODE $y' - y/x = -x^{-1} \cos(1/x)$. Find an initial condition for which the arbitrary constant becomes zero. Graph the resulting particular solution, experimenting to obtain a good figure near $x = 0$.

(b) Generalizing (a) from $n = 1$ to arbitrary n, solve the ODE $y' - ny/x = -x^{n-2} \cos(1/x)$. Find an initial condition as in (a) and experiment with the graph.

15–20 GENERAL PROPERTIES OF LINEAR ODEs

These properties are of practical and theoretical importance because they enable us to obtain new solutions from given ones. Thus in modeling, whenever possible, we prefer linear ODEs over nonlinear ones, which have no similar properties.

Show that nonhomogeneous linear ODEs (1) and homogeneous linear ODEs (2) have the following properties. Illustrate each property by a calculation for two or three equations of your choice. Give proofs.

15. The sum $y_1 + y_2$ of two solutions y_1 and y_2 of the homogeneous equation (2) is a solution of (2), and so is a scalar multiple ay_1 for any constant a. These properties are not true for (1)!

16. $y = 0$ (that is, $y(x) = 0$ for all x, also written $y(x) \equiv 0$) is a solution of (2) [not of (1) if $r(x) \neq 0$!], called the **trivial solution**.

17. The sum of a solution of (1) and a solution of (2) is a solution of (1).

18. The difference of two solutions of (1) is a solution of (2).

19. If y_1 is a solution of (1), what can you say about cy_1?

20. If y_1 and y_2 are solutions of $y_1' + py_1 = r_1$ and $y_2' + py_2 = r_2$, respectively (with the same p!), what can you say about the sum $y_1 + y_2$?

21. Variation of parameter. Another method of obtaining (4) results from the following idea. Write (3) as cy^*, where y^* is the exponential function, which is a solution of the homogeneous linear ODE $y^{*\prime} + py^* = 0$. Replace the arbitrary constant c in (3) with a function u to be determined so that the resulting function $y = uy^*$ is a solution of the nonhomogeneous linear ODE $y' + py = r$.

22–28 NONLINEAR ODEs

Using a method of this section or separating variables, find the general solution. If an initial condition is given, find also the particular solution and sketch or graph it.

22. $y' + y = y^2, \quad y(0) = -\frac{1}{3}$

23. $y' + xy = xy^{-1}, \quad y(0) = 3$

24. $y' + y = -x/y$

25. $y' = 3.2y - 10y^2$

26. $y' = (\tan y)/(x - 1), \quad y(0) = \frac{1}{2}\pi$

27. $y' = 1/(6e^y - 2x)$

28. $2xyy' + (x - 1)y^2 = x^2e^x \quad (\text{Set } y^2 = z)$

29. REPORT PROJECT. Transformation of ODEs. We have transformed ODEs to separable form, to exact form, and to linear form. The purpose of such transformations is an extension of solution methods to larger classes of ODEs. Describe the key idea of each of these transformations and give three typical examples of your choice for each transformation. Show each step (not just the transformed ODE).

30. TEAM PROJECT. Riccati Equation. Clairaut Equation. Singular Solution.

A **Riccati equation** is of the form

$$y' + p(x)y = g(x)y^2 + h(x). \tag{14}$$

A **Clairaut equation** is of the form

$$y = xy' + g(y'). \tag{15}$$

(a) Apply the transformation $y = Y + 1/u$ to the Riccati equation (14), where Y is a solution of (14), and obtain for u the linear ODE $u' + (2Yg - p)u = -g$. Explain the effect of the transformation by writing it as $y = Y + v$, $v = 1/u$.

(b) Show that $y = Y = x$ is a solution of the ODE $y' - (2x^3 + 1)\, y = -x^2y^2 - x^4 - x + 1$ and solve this Riccati equation, showing the details.

(c) Solve the Clairaut equation $y'^2 - xy' + y = 0$ as follows. Differentiate it with respect to x, obtaining $y''(2y' - x) = 0$. Then solve (A) $y'' = 0$ and (B) $2y' - x = 0$ separately and substitute the two solutions (a) and (b) of (A) and (B) into the given ODE. Thus obtain (a) a general solution (straight lines) and (b) a parabola for which those lines (a) are tangents (Fig. 6 in Prob. Set 1.1); so (b) is the envelope of (a). Such a solution (b) that cannot be obtained from a general solution is called a **singular solution**.

(d) Show that the Clairaut equation (15) has as solutions a family of straight lines $y = cx + g(c)$ and a singular solution determined by $g'(s) = -x$, where $s = y'$, that forms the envelope of that family.

31–40 MODELING. FURTHER APPLICATIONS

31. Newton's law of cooling. If the temperature of a cake is 300°F when it leaves the oven and is 200°F ten minutes later, when will it be practically equal to the room temperature of 60°F, say, when will it be 61°F?

32. Heating and cooling of a building. Heating and cooling of a building can be modeled by the ODE

$$T' = k_1(T - T_a) + k_2(T - T_\omega) + P,$$

where $T = T(t)$ is the temperature in the building at time t, T_a the outside temperature, T_w the temperature wanted in the building, and P the rate of increase of T due to machines and people in the building, and k_1 and k_2 are (negative) constants. Solve this ODE, assuming $P = \text{const}$, $T_w = \text{const}$, and T_a varying sinusoidally over 24 hours, say, $T_a = A - C\cos(2\pi/24)t$. Discuss the effect of each term of the equation on the solution.

33. Drug injection. Find and solve the model for drug injection into the bloodstream if, beginning at $t = 0$, a constant amount A g/min is injected and the drug is simultaneously removed at a rate proportional to the amount of the drug present at time t.

34. Epidemics. A model for the spread of contagious diseases is obtained by assuming that the rate of spread is proportional to the number of contacts between infected and noninfected persons, who are assumed to move freely among each other. Set up the model. Find the equilibrium solutions and indicate their stability or instability. Solve the ODE. Find the limit of the proportion of infected persons as $t \to \infty$ and explain what it means.

35. Lake Erie. Lake Erie has a water volume of about 450 km^3 and a flow rate (in and out) of about 175 km^2

per year. If at some instant the lake has pollution concentration $p = 0.04\%$, how long, approximately, will it take to decrease it to $p/2$, assuming that the inflow is much cleaner, say, it has pollution concentration $p/4$, and the mixture is uniform (an assumption that is only imperfectly true)? First guess.

36. Harvesting renewable resources. Fishing. Suppose that the population $y(t)$ of a certain kind of fish is given by the logistic equation (11), and fish are caught at a rate Hy proportional to y. Solve this so-called *Schaefer model.* Find the equilibrium solutions y_1 and y_2 (> 0) when $H < A$. The expression $Y = Hy_2$ is called the **equilibrium harvest** or **sustainable yield** *corresponding to H.* Why?

37. Harvesting. In Prob. 36 find and graph the solution satisfying $y(0) = 2$ when (for simplicity) $A = B = 1$ and $H = 0.2$. What is the limit? What does it mean? What if there were no fishing?

38. Intermittent harvesting. In Prob. 36 assume that you fish for 3 years, then fishing is banned for the next 3 years. Thereafter you start again. And so on. This is called *intermittent harvesting.* Describe qualitatively how the population will develop if intermitting is continued periodically. Find and graph the solution for the first 9 years, assuming that $A = B = 1$, $H = 0.2$, and $y(0) = 2$.

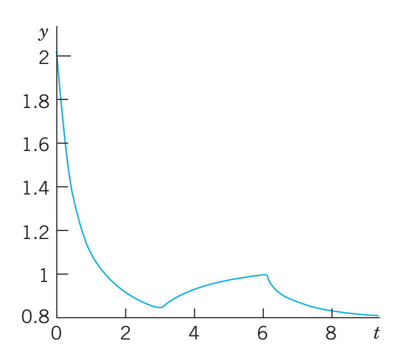


Fig. 23. Fish population in Problem 38

39. Extinction vs. unlimited growth. If in a population $y(t)$ the death rate is proportional to the population, and the birth rate is proportional to the chance encounters of meeting mates for reproduction, what will the model be? Without solving, find out what will eventually happen to a small initial population. To a large one. Then solve the model.

40. Air circulation. In a room containing 20,000 ft^3 of air, 600 ft^3 of fresh air flows in per minute, and the mixture (made practically uniform by circulating fans) is exhausted at a rate of 600 cubic feet per minute (cfm). What is the amount of fresh air $y(t)$ at any time if $y(0) = 0$? After what time will 90% of the air be fresh?

1.6 Orthogonal Trajectories. *Optional*

An important type of problem in physics or geometry is to find a family of curves that intersects a given family of curves at right angles. The new curves are called **orthogonal trajectories** of the given curves (and conversely). Examples are curves of equal temperature **(isotherms)** and curves of heat flow, curves of equal altitude **(contour lines)** on a map and curves of steepest descent on that map, curves of equal potential (**equipotential curves**, curves of equal voltage—the ellipses in Fig. 24) and curves of electric force (the parabolas in Fig. 24).

Here the **angle of intersection** between two curves is defined to be the angle between the tangents of the curves at the intersection point. *Orthogonal* is another word for *perpendicular.*

In many cases orthogonal trajectories can be found using ODEs. In general, if we consider $G(x, y, c) = 0$ to be a given family of curves in the xy-plane, then each value of c gives a particular curve. Since c is one parameter, such a family is called a **one-parameter family of curves**.

In detail, let us explain this method by a family of ellipses

$$\tfrac{1}{2}x^2 + y^2 = c \qquad (c > 0) \tag{1}$$

and illustrated in Fig. 24. We assume that this family of ellipses represents electric equipotential curves between the two black ellipses (equipotential surfaces between two elliptic cylinders in space, of which Fig. 24 shows a cross-section). We seek the orthogonal trajectories, the curves of electric force. Equation (1) is a *one-parameter family* with *parameter* c. Each value of c (> 0) corresponds to one of these ellipses.

Step 1. Find an ODE for which the given family is a general solution. Of course, this ODE must no longer contain the parameter c. Differentiating (1), we have $x + 2yy' = 0$. Hence the ODE of the given curves is

$$y' = f(x, y) = -\frac{x}{2y}. \tag{2}$$

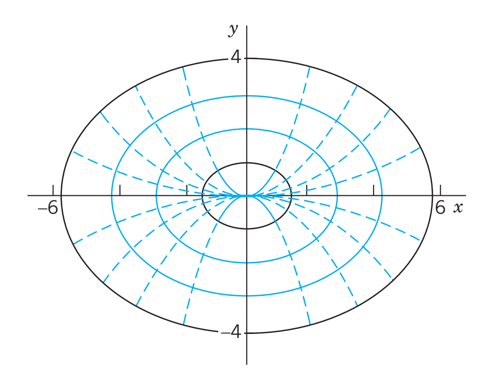


Fig. 24. Electrostatic field between two ellipses (elliptic cylinders in space): Elliptic equipotential curves (equipotential surfaces) and orthogonal trajectories (parabolas)

Step 2. Find an ODE for the orthogonal trajectories $\tilde{y} = \tilde{y}(x)$. This ODE is

$$\tilde{y}' = -\frac{1}{f(x, \tilde{y})} = +\frac{2\tilde{y}}{x} \tag{3}$$

with the same f as in (2). Why? Well, a given curve passing through a point (x_0, y_0) has slope $f(x_0, y_0)$ at that point, by (2). The trajectory through (x_0, y_0) has slope $-1/f(x_0, y_0)$ by (3). The product of these slopes is -1, as we see. From calculus it is known that this is the condition for orthogonality (perpendicularity) of two straight lines (the tangents at (x_0, y_0)), hence of the curve and its orthogonal trajectory at (x_0, y_0).

Step 3. Solve (3) by separating variables, integrating, and taking exponents:

$$\frac{d\tilde{y}}{\tilde{y}} = 2\frac{dx}{x}, \qquad \ln|\tilde{y}| = 2 \ln x + c, \qquad \tilde{y} = c^* x^2.$$

This is the family of orthogonal trajectories, the quadratic parabolas along which electrons or other charged particles (of very small mass) would move in the electric field between the black ellipses (elliptic cylinders).

PROBLEM SET 1.6

1–3 FAMILIES OF CURVES

Represent the given family of curves in the form $G(x, y; c) = 0$ and sketch some of the curves.

1. All ellipses with foci -3 and 3 on the x-axis.

2. All circles with centers on the cubic parabola $y = x^3$ and passing through the origin (0, 0).

3. The catenaries obtained by translating the catenary $y = \cosh x$ in the direction of the straight line $y = x$.

4–10 ORTHOGONAL TRAJECTORIES (OTs)

Sketch or graph some of the given curves. Guess what their OTs may look like. Find these OTs.

4. $y = x^2 + c$

5. $y = cx$

6. $xy = c$

7. $y = c/x^2$

8. $y = \sqrt{x + c}$

9. $y = ce^{-x^2}$

10. $x^2 + (y - c)^2 = c^2$

11–16 APPLICATIONS, EXTENSIONS

11. Electric field. Let the electric **equipotential lines** (curves of constant potential) between two concentric cylinders with the z-axis in space be given by $u(x, y) = x^2 + y^2 = c$ (these are circular cylinders in the xyz-space). Using the method in the text, find their orthogonal trajectories (the curves of electric force).

12. Electric field. The lines of electric force of two opposite charges of the same strength at $(-1, 0)$ and $(1, 0)$ are the circles through $(-1, 0)$ and $(1, 0)$. Show that these circles are given by $x^2 + (y - c)^2 = 1 + c^2$. Show that the **equipotential lines** (which are orthogonal trajectories of those circles) are the circles given by $(x + c^*)^2 + \tilde{y}^2 = c^{*2} - 1$ (dashed in Fig. 25).

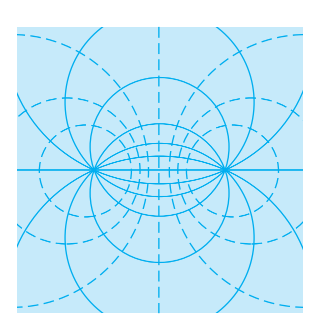


Fig. 25. Electric field in Problem 12

13. Temperature field. Let the **isotherms** (curves of constant temperature) in a body in the upper half-plane $y > 0$ be given by $4x^2 + 9y^2 = c$. Find the orthogonal trajectories (the curves along which heat will flow in regions filled with heat-conducting material and free of heat sources or heat sinks).

14. Conic sections. Find the conditions under which the orthogonal trajectories of families of ellipses $x^2/a^2 + y^2/b^2 = c$ are again conic sections. Illustrate your result graphically by sketches or by using your CAS. What happens if $a \to 0$? If $b \to 0$?

15. Cauchy–Riemann equations. Show that for a family $u(x, y) = c = \text{const}$ the orthogonal trajectories $v(x, y) = c^* = \text{const}$ can be obtained from the following *Cauchy–Riemann equations* (which are basic in complex analysis in Chap. 13) and use them to find the orthogonal trajectories of $e^x \sin y = \text{const}$. (Here, subscripts denote partial derivatives.)

$$u_x = v_y, \qquad u_y = -v_x$$

16. Congruent OTs. If $y' = f(x)$ with f independent of y, show that the curves of the corresponding family are congruent, and so are their OTs.

1.7 Existence and Uniqueness of Solutions for Initial Value Problems

The initial value problem

$$|y'| + |y| = 0, \qquad y(0) = 1$$

has no solution because $y = 0$ (that is, $y(x) = 0$ for all x) is the only solution of the ODE. The initial value problem

$$y' = 2x, \qquad y(0) = 1$$

has precisely one solution, namely, $y = x^2 + 1$. The initial value problem

$$xy' = y - 1, \qquad y(0) = 1$$

has infinitely many solutions, namely, $y = 1 + cx$, where c is an arbitrary constant because $y(0) = 1$ for all c.

From these examples we see that an **initial value problem**

$$\textbf{(1)} \qquad y' = f(x, y), \qquad y(x_0) = y_0$$

may have no solution, precisely one solution, or more than one solution. This fact leads to the following two fundamental questions.

Problem of Existence

Under what conditions does an initial value problem of the form (1) *have at least one solution (hence one or several solutions)?*

Problem of Uniqueness

Under what conditions does that problem have at most one solution (hence excluding the case that is has more than one solution)?

Theorems that state such conditions are called **existence theorems** and **uniqueness theorems**, respectively.

Of course, for our simple examples, we need no theorems because we can solve these examples by inspection; however, for complicated ODEs such theorems may be of considerable practical importance. Even when you are sure that your physical or other system behaves uniquely, occasionally your model may be oversimplified and may not give a faithful picture of reality.

THEOREM 1

Existence Theorem

Let the right side $f(x, y)$ of the ODE in the initial value problem

$$(1) \qquad y' = f(x, y), \qquad y(x_0) = y_0$$

be continuous at all points (x, y) in some rectangle

$$R: |x - x_0| < a, \qquad |y - y_0| < b \qquad \text{(Fig. 26)}$$

and **bounded** *in R; that is, there is a number K such that*

$$(2) \qquad |f(x, y)| \leqq K \qquad \textit{for all } (x, y) \textit{ in } R.$$

Then the initial value problem (1) *has at least one solution $y(x)$. This solution exists at least for all x in the subinterval $|x - x_0| < \alpha$ of the interval $|x - x_0| < a$; here, α is the smaller of the two numbers a and b/K.*

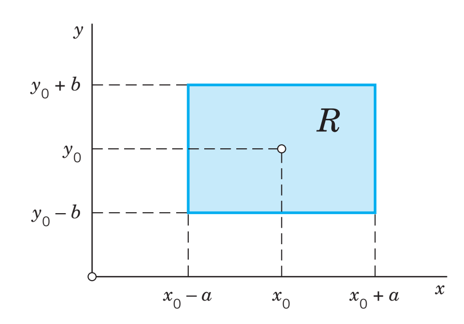


Fig. 26. Rectangle R in the existence and uniqueness theorems

(*Example of Boundedness.* The function $f(x, y) = x^2 + y^2$ is bounded (with $K = 2$) in the square $|x| < 1, |y| < 1$. The function $f(x, y) = \tan(x + y)$ is not bounded for $|x + y| < \pi/2$. Explain!)

THEOREM 2

Uniqueness Theorem

Let f and its partial derivative $f_y = \partial f/\partial y$ be continuous for all (x, y) in the rectangle R (Fig. 26) and bounded, say,

$$(3) \qquad \text{(a)} \quad |f(x, y)| \leqq K, \qquad \text{(b)} \quad |f_y(x, y)| \leqq M \qquad \textit{for all } (x, y) \textit{ in } R.$$

Then the initial value problem (1) has at most one solution $y(x)$. Thus, by Theorem 1, the problem has precisely one solution. This solution exists at least for all x in that subinterval $|x - x_0| < \alpha$.

Understanding These Theorems

These two theorems take care of almost all practical cases. Theorem 1 says that if $f(x, y)$ is continuous in some region in the xy-plane containing the point (x_0, y_0), then the initial value problem (1) has at least one solution.

Theorem 2 says that if, moreover, the partial derivative $\partial f/\partial y$ of f with respect to y exists and is continuous in that region, then (1) can have at most one solution; hence, by Theorem 1, it has precisely one solution.

Read again what you have just read—these are entirely new ideas in our discussion.

Proofs of these theorems are beyond the level of this book (see Ref. [A11] in App. 1); however, the following remarks and examples may help you to a good understanding of the theorems.

Since $y' = f(x, y)$, the condition (2) implies that $|y'| \leqq K$; that is, the slope of any solution curve $y(x)$ in R is at least $-K$ and at most K. Hence a solution curve that passes through the point (x_0, y_0) must lie in the colored region in Fig. 27 bounded by the lines l_1 and l_2 whose slopes are $-K$ and K, respectively. Depending on the form of R, two different cases may arise. In the first case, shown in Fig. 27a, we have $b/K \geqq a$ and therefore $\alpha = a$ in the existence theorem, which then asserts that the solution exists for all x between $x_0 - a$ and $x_0 + a$. In the second case, shown in Fig. 27b, we have $b/K < a$. Therefore, $\alpha = b/K < a$, and all we can conclude from the theorems is that the solution

exists for all x between $x_0 - b/K$ and $x_0 + b/K$. For larger or smaller x's the solution curve may leave the rectangle R, and since nothing is assumed about f outside R, nothing can be concluded about the solution for those larger or amaller x's; that is, for such x's the solution may or may not exist—we don't know.

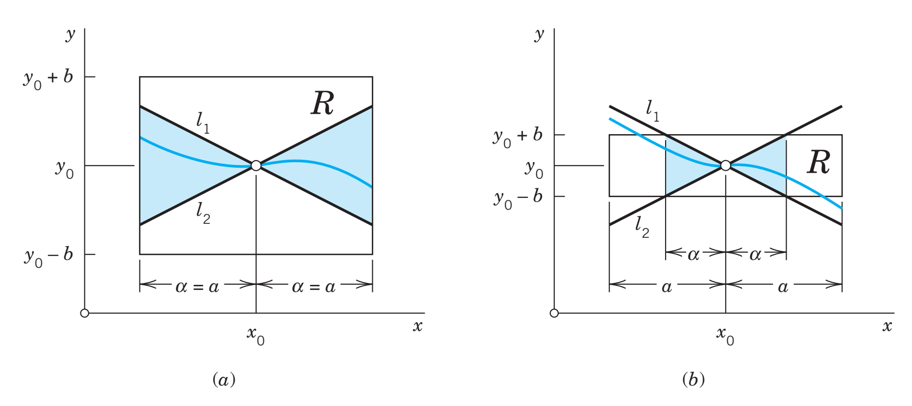


Fig. 27. The condition (2) of the existence theorem. (a) First case. (b) Second case

Let us illustrate our discussion with a simple example. We shall see that our choice of a rectangle R with a large base (a long x-interval) will lead to the case in Fig. 27b.

EXAMPLE 1 Choice of a Rectangle

Consider the initial value problem

$$y' = 1 + y^2, \qquad y(0) = 0$$

and take the rectangle R; $|x| < 5$, $|y| < 3$. Then $a = 5$, $b = 3$, and

$$|f(x, y)| = |1 + y^2| \leqq K = 10,$$

$$\left|\frac{\partial f}{\partial y}\right| = 2|y| \leqq M = 6,$$

$$\alpha = \frac{b}{K} = 0.3 < a.$$

Indeed, the solution of the problem is $y = \tan x$ (see Sec. 1.3, Example 1). This solution is discontinuous at $\pm\pi/2$, and there is no *continuous* solution valid in the entire interval $|x| < 5$ from which we started. ■

The conditions in the two theorems are sufficient conditions rather than necessary ones, and can be lessened. In particular, by the mean value theorem of differential calculus we have

$$f(x, y_2) - f(x, y_1) = (y_2 - y_1) \left.\frac{\partial f}{\partial y}\right|_{y=\tilde{y}}$$

where (x, y_1) and (x, y_2) are assumed to be in R, and $\tilde{y}$ is a suitable value between y_1 and y_2. From this and (3b) it follows that

$$|f(x, y_2) - f(x, y_1)| \leqq M|y_2 - y_1|. \tag{4}$$

It can be shown that (3b) may be replaced by the weaker condition (4), which is known as a **Lipschitz condition**.[9] However, continuity of $f(x, y)$ is not enough to guarantee the *uniqueness* of the solution. This may be illustrated by the following example.

EXAMPLE 2 Nonuniqueness

The initial value problem

$$y' = \sqrt{|y|}. \qquad y(0) = 0$$

has the two solutions

$$y = 0 \qquad \text{and} \qquad y^* = \begin{cases} x^2/4 & \text{if} \quad x \geqq 0 \\ -x^2/4 & \text{if} \quad x < 0 \end{cases}$$

although $f(x, y) = \sqrt{|y|}$ is continuous for all y. The Lipschitz condition (4) is violated in any region that includes the line $y = 0$, because for $y_1 = 0$ and positive y_2 we have

$$\text{(5)} \qquad \frac{|f(x, y_2) - f(x, y_1)|}{|y_2 - y_1|} = \frac{\sqrt{y_2}}{y_2} = \frac{1}{\sqrt{y_2}}, \qquad (\sqrt{y_2} > 0)$$

and this can be made as large as we please by choosing y_2 sufficiently small, whereas (4) requires that the quotient on the left side of (5) should not exceed a fixed constant M. ■

PROBLEM SET 1.7

1. **Linear ODE.** If p and r in $y' + p(x)y = r(x)$ are continuous for all x in an interval $|x - x_0| \leq a$, show that $f(x, y)$ in this ODE satisfies the conditions of our present theorems, so that a corresponding initial value problem has a unique solution. Do you actually need these theorems for this ODE?
2. **Existence?** Does the initial value problem $(x - 2)y' = y$, $y(2) = 1$ have a solution? Does your result contradict our present theorems?
3. **Vertical strip.** If the assumptions of Theorems 1 and 2 are satisfied not merely in a rectangle but in a vertical infinite strip $|x - x_0| < a$, in what interval will the solution of (1) exist?
4. **Change of initial condition.** What happens in Prob. 2 if you replace $y(2) = 1$ with $y(2) = k$?
5. **Length of x-interval.** In most cases the solution of an initial value problem (1) exists in an x-interval larger than that guaranteed by the present theorems. Show this fact for $y' = 2y^2$, $y(1) = 1$ by finding the best possible a (choosing b optimally) and comparing the result with the actual solution.
6. **CAS PROJECT. Picard Iteration. (a)** Show that by integrating the ODE in (1) and observing the initial condition you obtain

$$\text{(6)} \qquad y(x) = y_0 + \int_{x_0}^{x} f(t, y(t))\,dt.$$

This form (6) of (1) suggests **Picard's Iteration Method**[10] which is defined by

$$\text{(7)} \quad y_n(x) = y_0 + \int_{x_0}^{x} f(t, y_{n-1}(t)\,dt, \quad n = 1, 2, \cdots.$$

It gives approximations $y_1, y_2, y_3, \ldots$ of the unknown solution y of (1). Indeed, you obtain y_1 by substituting $y = y_0$ on the right and integrating—this is the first step—then y_2 by substituting $y = y_1$ on the right and integrating—this is the second step—and so on. Write

[9] RUDOLF LIPSCHITZ (1832–1903), German mathematician. Lipschitz and similar conditions are important in modern theories, for instance, in partial differential equations.

[10] EMILE PICARD (1856–1941). French mathematician, also known for his important contributions to complex analysis (see Sec. 16.2 for his famous theorem). Picard used his method to prove Theorems 1 and 2 as well as the convergence of the sequence (7) to the solution of (1). In precomputer times, the iteration was of little *practical* value because of the integrations.

a program of the iteration that gives a printout of the first approximations $y_0, y_1, \dots, y_N$ as well as their graphs on common axes. Try your program on two initial value problems of your own choice.

(b) Apply the iteration to $y' = x + y$, $y(0) = 0$. Also solve the problem exactly.

(c) Apply the iteration to $y' = 2y^2$, $y(0) = 1$. Also solve the problem exactly.

(d) Find all solutions of $y' = 2\sqrt{y}$, $y(1) = 0$. Which of them does Picard's iteration approximate?

(e) Experiment with the conjecture that Picard's iteration converges to the solution of the problem for any initial choice of y in the integrand in (7) (leaving y_0 outside the integral as it is). Begin with a simple ODE and see what happens. When you are reasonably sure, take a slightly more complicated ODE and give it a try.

7. Maximum α. What is the largest possible α in Example 1 in the text?

8. Lipschitz condition. Show that for a linear ODE $y' + p(x)y = r(x)$ with continuous p and r in $|x - x_0| \leqq a$ a Lipschitz condition holds. This is remarkable because it means that for a *linear* ODE the continuity of $f(x, y)$ guarantees not only the existence but also the uniqueness of the solution of an initial value problem. (Of course, this also follows directly from (4) in Sec. 1.5.)

9. Common points. Can two solution curves of the same ODE have a common point in a rectangle in which the assumptions of the present theorems are satisfied?

10. Three possible cases. Find all initial conditions such that $(x^2 - x)y' = (2x - 1)y$ has no solution, precisely one solution, and more than one solution.

CHAPTER 1 REVIEW QUESTIONS AND PROBLEMS

1. Explain the basic concepts ordinary and partial differential equations (ODEs, PDEs), order, general and particular solutions, initial value problems (IVPs). Give examples.
2. What is a linear ODE? Why is it easier to solve than a nonlinear ODE?
3. Does every first-order ODE have a solution? A solution formula? Give examples.
4. What is a direction field? A numeric method for first-order ODEs?
5. What is an exact ODE? Is $f(x)\,dx + g(y)\,dy = 0$ always exact?
6. Explain the idea of an integrating factor. Give two examples.
7. What other solution methods did we consider in this chapter?
8. Can an ODE sometimes be solved by several methods? Give three examples.
9. What does modeling mean? Can a CAS solve a model given by a first-order ODE? Can a CAS set up a model?
10. Give problems from mechanics, heat conduction, and population dynamics that can be modeled by first-order ODEs.

11–16 DIRECTION FIELD: NUMERIC SOLUTION

Graph a direction field (by a CAS or by hand) and sketch some solution curves. Solve the ODE exactly and compare. In Prob. 16 use Euler's method.

11. $y' + 2y = 0$
12. $y' = 1 - y^2$
13. $y' = y - 4y^2$
14. $xy' = y + x^2$
15. $y' + y = 1.01 \cos 10x$
16. Solve $y' = y - y^2$, $y(0) = 0.2$ by Euler's method (10 steps, $h = 0.1$). Solve exactly and compute the error.

17–21 GENERAL SOLUTION

Find the general solution. Indicate which method in this chapter you are using. Show the details of your work.

17. $y' + 2.5y = 1.6x$
18. $y' - 0.4y = 29 \sin x$
19. $25yy' - 4x = 0$
20. $y' = ay + by^2 \quad (a \neq 0)$
21. $(3xe^y + 2y)\,dx + (x^2e^y + x)\,dy = 0$

22–26 INITIAL VALUE PROBLEM (IVP)

Solve the IVP. Indicate the method used. Show the details of your work.

22. $y' + 4xy = e^{-2x^2}, \quad y(0) = -4.3$
23. $y' = \sqrt{1 - y^2}, \quad y(0) = 1/\sqrt{2}$
24. $y' + \frac{1}{2}y = y^3, \quad y(0) = \frac{1}{3}$
25. $3 \sec y\,dx + \frac{1}{3} \sec x\,dy = 0, \quad y(0) = 0$
26. $x \sinh y\,dy = \cosh y\,dx, \quad y(3) = 0$

27–30 MODELING, APPLICATIONS

27. **Exponential growth.** If the growth rate of a culture of bacteria is proportional to the number of bacteria present and after 1 day is 1.25 times the original number, within what interval of time will the number of bacteria (a) double, (b) triple?

28. Mixing problem. The tank in Fig. 28 contains 80 lb of salt dissolved in 500 gal of water. The inflow per minute is 20 lb of salt dissolved in 20 gal of water. The outflow is 20 gal/min of the uniform mixture. Find the time when the salt content $y(t)$ in the tank reaches 95% of its limiting value (as $t \to \infty$).

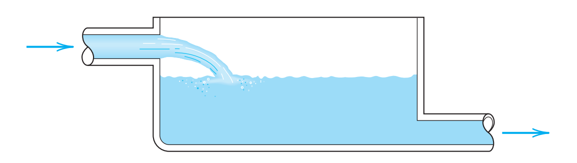


Fig. 28. Tank in Problem 28

29. Half-life. If in a reactor, uranium $^{237}_{97}\text{U}$ loses 10% of its weight within one day, what is its half-life? How long would it take for 99% of the original amount to disappear?

30. Newton's law of cooling. A metal bar whose temperature is 20°C is placed in boiling water. How long does it take to heat the bar to practically 100°C, say, to 99.9°C, if the temperature of the bar after 1 min of heating is 51.5°C? First guess, then calculate.

SUMMARY OF CHAPTER 1
First-Order ODEs

This chapter concerns **ordinary differential equations (ODEs) of first order** and their applications. These are equations of the form

$$(1) \qquad F(x, y, y') = 0 \qquad \text{or in explicit form} \qquad y' = f(x, y)$$

involving the derivative $y' = dy/dx$ of an unknown function y, given functions of x, and, perhaps, y itself. If the independent variable x is time, we denote it by t.

In Sec. 1.1 we explained the basic concepts and the process of **modeling**, that is, of expressing a physical or other problem in some mathematical form and solving it. Then we discussed the method of direction fields (Sec. 1.2), solution methods and models (Secs. 1.3–1.6), and, finally, ideas on existence and uniqueness of solutions (Sec. 1.7).

A first-order ODE usually has a **general solution**, that is, a solution involving an arbitrary constant, which we denote by c. In applications we usually have to find a unique solution by determining a value of c from an **initial condition** $y(x_0) = y_0$. Together with the ODE this is called an **initial value problem**

$$(2) \qquad y' = f(x, y), \qquad y(x_0) = y_0 \qquad (x_0, y_0 \text{ given numbers})$$

and its solution is a **particular solution** of the ODE. Geometrically, a general solution represents a family of curves, which can be graphed by using **direction fields** (Sec. 1.2). And each particular solution corresponds to one of these curves.

A **separable ODE** is one that we can put into the form

$$(3) \qquad g(y)\, dy = f(x)\, dx \qquad \text{(Sec. 1.3)}$$

by algebraic manipulations (possibly combined with transformations, such as $y/x = u$) and solve by integrating on both sides.

An **exact ODE** is of the form

(4) $$M(x, y)\,dx + N(x, y)\,dy = 0 \qquad \text{(Sec. 1.4)}$$

where $M\,dx + N\,dy$ is the **differential**

$$du = u_x\,dx + u_y\,dy$$

of a function $u(x, y)$, so that from $du = 0$ we immediately get the implicit general solution $u(x, y) = c$. This method extends to nonexact ODEs that can be made exact by multiplying them by some function $F(x, y,)$, called an **integrating factor** (Sec. 1.4).

Linear ODEs

(5) $$y' + p(x)y = r(x)$$

are very important. Their solutions are given by the integral formula (4), Sec. 1.5. Certain nonlinear ODEs can be transformed to linear form in terms of new variables. This holds for the **Bernoulli equation**

$$y' + p(x)y = g(x)y^a \qquad \text{(Sec. 1.5).}$$

Applications and ***modeling*** are discussed throughout the chapter, in particular in Secs. 1.1, 1.3, 1.5 (***population dynamics***, etc.), and 1.6 (***trajectories***).

Picard's ***existence*** and ***uniqueness theorems*** are explained in Sec. 1.7 (and *Picard's iteration* in Problem Set 1.7).

Numeric methods for first-order ODEs can be studied in Secs. 21.1 and 21.2 immediately after this chapter, as indicated in the chapter opening.

CHAPTER 2

Second-Order Linear ODEs

Many important applications in mechanical and electrical engineering, as shown in Secs. 2.4, 2.8, and 2.9, are modeled by linear ordinary differential equations (linear ODEs) of the second order. Their theory is representative of all linear ODEs as is seen when compared to linear ODEs of third and higher order, respectively. However, the solution formulas for second-order linear ODEs are simpler than those of higher order, so it is a natural progression to study ODEs of second order first in this chapter and then of higher order in Chap. 3.

Although ordinary differential equations (ODEs) can be grouped into linear and nonlinear ODEs, nonlinear ODEs are difficult to solve in contrast to linear ODEs for which many beautiful standard methods exist.

Chapter 2 includes the derivation of general and particular solutions, the latter in connection with initial value problems.

For those interested in solution methods for Legendre's, Bessel's, and the hypergeometric equations consult Chap. 5 and for Sturm–Liouville problems Chap. 11.

COMMENT. ***Numerics for second-order ODEs can be studied immediately after this chapter.*** See Sec. 21.3, which is independent of other sections in Chaps. 19–21.

Prerequisite: Chap. 1, in particular, Sec. 1.5.
Sections that may be omitted in a shorter course: 2.3, 2.9, 2.10.
References and Answers to Problems: App. 1 Part A, and App. 2.

2.1 Homogeneous Linear ODEs of Second Order

We have already considered first-order linear ODEs (Sec. 1.5) and shall now define and discuss linear ODEs of second order. These equations have important engineering applications, especially in connection with mechanical and electrical vibrations (Secs. 2.4, 2.8, 2.9) as well as in wave motion, heat conduction, and other parts of physics, as we shall see in Chap. 12.

A second-order ODE is called **linear** if it can be written

$$y'' + p(x)y' + q(x)y = r(x) \tag{1}$$

and **nonlinear** if it cannot be written in this form.

The distinctive feature of this equation is that it is *linear in y and its derivatives,* whereas the functions p, q, and r on the right may be any given functions of x. If the equation begins with, say, $f(x)y''$, then divide by $f(x)$ to have the **standard form** (1) with y'' as the first term.

The definitions of homogeneous and nonhomogenous second-order linear ODEs are very similar to those of first-order ODEs discussed in Sec. 1.5. Indeed, if $r(x) \equiv 0$ (that is, $r(x) = 0$ for all x considered; read "$r(x)$ is identically zero"), then (1) reduces to

$$y'' + p(x)y' + q(x)y = 0 \tag{2}$$

and is called **homogeneous**. If $r(x) \not\equiv 0$, then (1) is called **nonhomogeneous**. This is similar to Sec. 1.5.

An example of a nonhomogeneous linear ODE is

$$y'' + 25y = e^{-x} \cos x,$$

and a homogeneous linear ODE is

$$xy'' + y' + xy = 0, \qquad \text{written in standard form} \qquad y'' + \frac{1}{x}y' + y = 0.$$

Finally, an example of a nonlinear ODE is

$$y''y + y'^2 = 0.$$

The functions p and q in (1) and (2) are called the **coefficients** of the ODEs.

Solutions are defined similarly as for first-order ODEs in Chap. 1. A function

$$y = h(x)$$

is called a *solution* of a (linear or nonlinear) second-order ODE on some open interval I if h is defined and twice differentiable throughout that interval and is such that the ODE becomes an identity if we replace the unknown y by h, the derivative y' by h', and the second derivative y'' by h''. Examples are given below.

Homogeneous Linear ODEs: Superposition Principle

Sections 2.1–2.6 will be devoted to **homogeneous** linear ODEs (2) and the remaining sections of the chapter to nonhomogeneous linear ODEs.

Linear ODEs have a rich solution structure. For the homogeneous equation the backbone of this structure is the *superposition principle* or *linearity principle,* which says that we can obtain further solutions from given ones by adding them or by multiplying them with any constants. Of course, this is a great advantage of homogeneous linear ODEs. Let us first discuss an example.

EXAMPLE 1 Homogeneous Linear ODEs: Superposition of Solutions

The functions $y = \cos x$ and $y = \sin x$ are solutions of the homogeneous linear ODE

$$y'' + y = 0$$

for all x. We verify this by differentiation and substitution. We obtain $(\cos x)'' = -\cos x$; hence

$$y'' + y = (\cos x)'' + \cos x = -\cos x + \cos x = 0.$$

Similarly for $y = \sin x$ (verify!). We can go an important step further. We multiply $\cos x$ by any constant, for instance, 4.7, and $\sin x$ by, say, -2, and take the sum of the results, claiming that it is a solution. Indeed, differentiation and substitution gives

$$(4.7 \cos x - 2 \sin x)'' + (4.7 \cos x - 2 \sin x) = -4.7 \cos x + 2 \sin x + 4.7 \cos x - 2 \sin x = 0. \quad \blacksquare$$

In this example we have obtained from y_1 ($= \cos x$) and y_2 ($= \sin x$) a function of the form

$$y = c_1 y_1 + c_2 y_2 \qquad (c_1, c_2 \text{ arbitrary constants}). \tag{3}$$

This is called a **linear combination** of y_1 and y_2. In terms of this concept we can now formulate the result suggested by our example, often called the **superposition principle** or **linearity principle**.

THEOREM 1

Fundamental Theorem for the Homogeneous Linear ODE (2)

For a homogeneous linear ODE (2), *any linear combination of two solutions on an open interval I is again a solution of* (2) *on I. In particular, for such an equation, sums and constant multiples of solutions are again solutions.*

PROOF Let y_1 and y_2 be solutions of (2) on I. Then by substituting $y = c_1 y_1 + c_2 y_2$ and its derivatives into (2), and using the familiar rule $(c_1 y_1 + c_2 y_2)' = c_1 y_1' + c_2 y_2'$, etc., we get

$$\begin{aligned} y'' + py' + qy &= (c_1 y_1 + c_2 y_2)'' + p(c_1 y_1 + c_2 y_2)' + q(c_1 y_1 + c_2 y_2) \\ &= c_1 y_1'' + c_2 y_2'' + p(c_1 y_1' + c_2 y_2') + q(c_1 y_1 + c_2 y_2) \\ &= c_1(y_1'' + py_1' + qy_1) + c_2(y_2'' + py_2' + qy_2) = 0, \end{aligned}$$

since in the last line, $(\cdots) = 0$ because y_1 and y_2 are solutions, by assumption. This shows that y is a solution of (2) on I. ■

CAUTION! Don't forget that this highly important theorem holds for *homogeneous linear* ODEs only but ***does not hold*** for nonhomogeneous linear or nonlinear ODEs, as the following two examples illustrate.

EXAMPLE 2 **A Nonhomogeneous Linear ODE**

Verify by substitution that the functions $y = 1 + \cos x$ and $y = 1 + \sin x$ are solutions of the nonhomogeneous linear ODE

$$y'' + y = 1,$$

but their sum is not a solution. Neither is, for instance, $2(1 + \cos x)$ or $5(1 + \sin x)$. ■

EXAMPLE 3 **A Nonlinear ODE**

Verify by substitution that the functions $y = x^2$ and $y = 1$ are solutions of the nonlinear ODE

$$y''y - xy' = 0,$$

but their sum is not a solution. Neither is $-x^2$, so you cannot even multiply by -1! ■

Initial Value Problem. Basis. General Solution

Recall from Chap. 1 that for a first-order ODE, an *initial value problem* consists of the ODE and one *initial condition* $y(x_0) = y_0$. The initial condition is used to determine the *arbitrary constant c* in the *general solution* of the ODE. This results in a unique solution, as we need it in most applications. That solution is called a *particular solution* of the ODE. These ideas extend to second-order ODEs as follows.

For a second-order homogeneous linear ODE (2) an **initial value problem** consists of (2) and two **initial conditions**

$$y(x_0) = K_0, \qquad y'(x_0) = K_1. \tag{4}$$

These conditions prescribe given values K_0 and K_1 of the solution and its first derivative (the slope of its curve) at the same given $x = x_0$ in the open interval considered.

The conditions (4) are used to determine the two arbitrary constants c_1 and c_2 in a **general solution**

$$y = c_1 y_1 + c_2 y_2 \tag{5}$$

of the ODE; here, y_1 and y_2 are suitable solutions of the ODE, with "suitable" to be explained after the next example. This results in a unique solution, passing through the point (x_0, K_0) with K_1 as the tangent direction (the slope) at that point. That solution is called a **particular solution** of the ODE (2).

EXAMPLE 4 Initial Value Problem

Solve the initial value problem

$$y'' + y = 0, \qquad y(0) = 3.0, \qquad y'(0) = -0.5.$$

Solution. ***Step 1. General solution.*** The functions $\cos x$ and $\sin x$ are solutions of the ODE (by Example 1), and we take

$$y = c_1 \cos x + c_2 \sin x.$$

This will turn out to be a general solution as defined below.

Step 2. Particular solution. We need the derivative $y' = -c_1 \sin x + c_2 \cos x$. From this and the initial values we obtain, since $\cos 0 = 1$ and $\sin 0 = 0$,

$$y(0) = c_1 = 3.0 \qquad \text{and} \qquad y'(0) = c_2 = -0.5.$$

This gives as the solution of our initial value problem the particular solution

$$y = 3.0 \cos x - 0.5 \sin x.$$

Figure 29 shows that at $x = 0$ it has the value 3.0 and the slope -0.5, so that its tangent intersects the x-axis at $x = 3.0/0.5 = 6.0$. (The scales on the axes differ!) ■

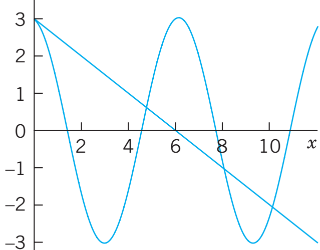


Fig. 29. Particular solution and initial tangent in Example 4

Observation. Our choice of y_1 and y_2 was general enough to satisfy both initial conditions. Now let us take instead two proportional solutions $y_1 = \cos x$ and $y_2 = k \cos x$, so that $y_1/y_2 = 1/k = \text{const}$. Then we can write $y = c_1 y_1 + c_2 y_2$ in the form

$$y = c_1 \cos x + c_2(k \cos x) = C \cos x \qquad \text{where} \qquad C = c_1 + c_2 k.$$

Hence we are no longer able to satisfy two initial conditions with only one arbitrary constant C. Consequently, in defining the concept of a general solution, we must exclude proportionality. And we see at the same time why the concept of a general solution is of importance in connection with initial value problems.

DEFINITION

General Solution, Basis, Particular Solution

A **general solution** of an ODE (2) on an open interval I is a solution (5) in which y_1 and y_2 are solutions of (2) on I that are not proportional, and c_1 and c_2 are arbitrary constants. These y_1, y_2 are called a **basis** (or a **fundamental system**) of solutions of (2) on I.

A **particular solution** of (2) on I is obtained if we assign specific values to c_1 and c_2 in (5).

For the definition of an *interval* see Sec. 1.1. Furthermore, as usual, y_1 and y_2 are called *proportional* on I if for all x on I,

$$(6) \qquad \text{(a)} \quad y_1 = ky_2 \qquad \text{or} \qquad \text{(b)} \quad y_2 = ly_1$$

where k and l are numbers, zero or not. (Note that (a) implies (b) if and only if $k \neq 0$).

Actually, we can reformulate our definition of a basis by using a concept of general importance. Namely, two functions y_1 and y_2 are called **linearly independent** on an interval I where they are defined if

$$(7) \qquad k_1y_1(x) + k_2y_2(x) = 0 \qquad \text{everywhere on } I \text{ implies} \qquad k_1 = 0 \text{ and } k_2 = 0.$$

And y_1 and y_2 are called **linearly dependent** on I if (7) also holds for some constants k_1, k_2 not both zero. Then, if $k_1 \neq 0$ or $k_2 \neq 0$, we can divide and see that y_1 and y_2 are proportional,

$$y_1 = -\frac{k_2}{k_1}y_2 \qquad \text{or} \qquad y_2 = -\frac{k_1}{k_2}y_1.$$

In contrast, in the case of linear *independence* these functions are not proportional because then we cannot divide in (7). This gives the following

DEFINITION

Basis (Reformulated)

A **basis** of solutions of (2) on an open interval I is a pair of linearly independent solutions of (2) on I.

If the coefficients p and q of (2) are continuous on some open interval I, then (2) has a general solution. It yields the unique solution of any initial value problem (2), (4). It includes all solutions of (2) on I; hence (2) has no *singular solutions* (solutions not obtainable from of a general solution; see also Problem Set 1.1). All this will be shown in Sec. 2.6.

EXAMPLE 5 **Basis, General Solution, Particular Solution**

$\cos x$ and $\sin x$ in Example 4 form a basis of solutions of the ODE $y'' + y = 0$ for all x because their quotient is $\cot x \neq$ const (or $\tan x \neq$ const). Hence $y = c_1 \cos x + c_2 \sin x$ is a general solution. The solution $y = 3.0 \cos x - 0.5 \sin x$ of the initial value problem is a particular solution. ■

EXAMPLE 6 **Basis, General Solution, Particular Solution**

Verify by substitution that $y_1 = e^x$ and $y_2 = e^{-x}$ are solutions of the ODE $y'' - y = 0$. Then solve the initial value problem

$$y'' - y = 0, \qquad y(0) = 6, \qquad y'(0) = -2.$$

Solution. $(e^x)'' - e^x = 0$ and $(e^{-x})'' - e^{-x} = 0$ show that e^x and e^{-x} are solutions. They are not proportional, $e^x/e^{-x} = e^{2x} \neq$ const. Hence e^x, e^{-x} form a basis for all x. We now write down the corresponding general solution and its derivative and equate their values at 0 to the given initial conditions,

$$y = c_1 e^x + c_2 e^{-x}, \qquad y' = c_1 e^x - c_2 e^{-x}, \qquad y(0) = c_1 + c_2 = 6, \qquad y'(0) = c_1 - c_2 = -2.$$

By addition and subtraction, $c_1 = 2$, $c_2 = 4$, so that the *answer* is $y = 2e^x + 4e^{-x}$. This is the particular solution satisfying the two initial conditions. ■

Find a Basis if One Solution Is Known. Reduction of Order

It happens quite often that one solution can be found by inspection or in some other way. Then a second linearly independent solution can be obtained by solving a first-order ODE. This is called the method of **reduction of order**.[1] We first show how this method works in an example and then in general.

EXAMPLE 7 **Reduction of Order if a Solution Is Known. Basis**

Find a basis of solutions of the ODE

$$(x^2 - x)y'' - xy' + y = 0.$$

Solution. Inspection shows that $y_1 = x$ is a solution because $y_1' = 1$ and $y_1'' = 0$, so that the first term vanishes identically and the second and third terms cancel. The idea of the method is to substitute

$$y = uy_1 = ux, \qquad y' = u'x + u, \qquad y'' = u''x + 2u'$$

into the ODE. This gives

$$(x^2 - x)(u''x + 2u') - x(u'x + u) + ux = 0.$$

ux and $-xu$ cancel and we are left with the following ODE, which we divide by x, order, and simplify,

$$(x^2 - x)(u''x + 2u') - x^2u' = 0, \qquad (x^2 - x)u'' + (x - 2)u' = 0.$$

This ODE is of first order in $v = u'$, namely, $(x^2 - x)v' + (x - 2)v = 0$. Separation of variables and integration gives

$$\frac{dv}{v} = -\frac{x - 2}{x^2 - x}\,dx = \left(\frac{1}{x - 1} - \frac{2}{x}\right)dx, \qquad \ln|v| = \ln|x - 1| - 2\ln|x| = \ln\frac{|x - 1|}{x^2}.$$

---

[1]Credited to the great mathematician JOSEPH LOUIS LAGRANGE (1736–1813), who was born in Turin, of French extraction, got his first professorship when he was 19 (at the Military Academy of Turin), became director of the mathematical section of the Berlin Academy in 1766, and moved to Paris in 1787. His important major work was in the calculus of variations, celestial mechanics, general mechanics (*Mécanique analytique,* Paris, 1788), differential equations, approximation theory, algebra, and number theory.

We need no constant of integration because we want to obtain a particular solution; similarly in the next integration. Taking exponents and integrating again, we obtain

$$v = \frac{x-1}{x^2} = \frac{1}{x} - \frac{1}{x^2}, \qquad u = \int v\,dx = \ln|x| + \frac{1}{x}, \qquad \text{hence} \qquad y_2 = ux = x \ln|x| + 1.$$

Since $y_1 = x$ and $y_2 = x \ln|x| + 1$ are linearly independent (their quotient is not constant), we have obtained a basis of solutions, valid for all positive x. ■

In this example we applied **reduction of order** to a homogeneous linear ODE [see (2)]

$$y'' + p(x)y' + q(x)y = 0.$$

Note that we now take the ODE in standard form, with y'', not $f(x)y''$—this is essential in applying our subsequent formulas. We assume a solution y_1 of (2), on an open interval I, to be known and want to find a basis. For this we need a second linearly independent solution y_2 of (2) on I. To get y_2, we substitute

$$y = y_2 = uy_1, \qquad y' = y_2' = u'y_1 + uy_1', \qquad y'' = y_2'' = u''y_1 + 2u'y_1' + uy_1''$$

into (2). This gives

$$u''y_1 + 2u'y_1' + uy_1'' + p(u'y_1 + uy_1') + quy_1 = 0. \tag{8}$$

Collecting terms in u'', u', and u, we have

$$u''y_1 + u'(2y_1' + py_1) + u(y_1'' + py_1' + qy_1) = 0.$$

Now comes the main point. Since y_1 is a solution of (2), the expression in the last parentheses is zero. Hence u is gone, and we are left with an ODE in u' and u''. We divide this remaining ODE by y_1 and set $u' = U$, $u'' = U'$,

$$u'' + u' \frac{2y_1' + py_1}{y_1} = 0, \qquad \text{thus} \qquad U' + \left(\frac{2y_1'}{y_1} + p\right)U = 0.$$

This is the desired first-order ODE, the reduced ODE. Separation of variables and integration gives

$$\frac{dU}{U} = -\left(\frac{2y_1'}{y_1} + p\right)dx \qquad \text{and} \qquad \ln|U| = -2\ln|y_1| - \int p\,dx.$$

By taking exponents we finally obtain

$$U = \frac{1}{y_1^2} e^{-\int p\,dx}. \tag{9}$$

Here $U = u'$, so that $u = \int U\,dx$. Hence the desired second solution is

$$y_2 = y_1 u = y_1 \int U\,dx.$$

The quotient $y_2/y_1 = u = \int U\,dx$ cannot be constant (since $U > 0$), so that y_1 and y_2 form a basis of solutions.

PROBLEM SET 2.1

REDUCTION OF ORDER is important because it gives a simpler ODE. A general second-order ODE $F(x, y, y', y'') = 0$, linear or not, can be reduced to first order if y does not occur explicitly (Prob. 1) or if x does not occur explicitly (Prob. 2) or if the ODE is homogeneous linear and we know a solution (see the text).

1. **Reduction.** Show that $F(x, y', y'') = 0$ can be reduced to first order in $z = y'$ (from which y follows by integration). Give two examples of your own.
2. **Reduction.** Show that $F(y, y', y'') = 0$ can be reduced to a first-order ODE with y as the independent variable and $y'' = (dz/dy)z$, where $z = y'$; derive this by the chain rule. Give two examples.

3–10 REDUCTION OF ORDER

Reduce to first order and solve, showing each step in detail.

3. $y'' + y' = 0$
4. $2xy'' = 3y'$
5. $yy'' = 3y'^2$
6. $xy'' + 2y' + xy = 0, \quad y_1 = (\cos x)/x$
7. $y'' + y'^3 \sin y = 0$
8. $y'' = 1 + y'^2$
9. $x^2y'' - 5xy' + 9y = 0, \quad y_1 = x^3$
10. $y'' + (1 + 1/y)y'^2 = 0$

11–14 APPLICATIONS OF REDUCIBLE ODEs

11. **Curve.** Find the curve through the origin in the xy-plane which satisfies $y'' = 2y'$ and whose tangent at the origin has slope 1.
12. **Hanging cable.** It can be shown that the curve $y(x)$ of an inextensible flexible homogeneous cable hanging between two fixed points is obtained by solving $y'' = k\sqrt{1 + y'^2}$, where the constant k depends on the weight. This curve is called *catenary* (from Latin *catena* = the chain). Find and graph $y(x)$, assuming that $k = 1$ and those fixed points are $(-1, 0)$ and $(1, 0)$ in a vertical xy-plane.
13. **Motion.** If, in the motion of a small body on a straight line, the sum of velocity and acceleration equals a positive constant, how will the distance $y(t)$ depend on the initial velocity and position?
14. **Motion.** In a straight-line motion, let the velocity be the reciprocal of the acceleration. Find the distance $y(t)$ for arbitrary initial position and velocity.

15–19 GENERAL SOLUTION. INITIAL VALUE PROBLEM (IVP)

(More in the next set.) **(a)** Verify that the given functions are linearly independent and form a basis of solutions of the given ODE. **(b)** Solve the IVP. Graph or sketch the solution.

15. $4y'' + 25y = 0, \quad y(0) = 3.0, \quad y'(0) = -2.5,$ $\cos 2.5x, \sin 2.5x$
16. $y'' + 0.6y' + 0.09y = 0, \quad y(0) = 2.2, \ y'(0) = 0.14,$ $e^{-0.3x}, xe^{-0.3x}$
17. $4x^2y'' - 3y = 0, \quad y(1) = -3, \quad y'(1) = 0,$ $x^{3/2}, x^{-1/2}$
18. $x^2y'' - xy' + y = 0, \quad y(1) = 4.3, \quad y'(1) = 0.5,$ $x, x \ln x$
19. $y'' + 2y' + 2y = 0, \quad y(0) = 0, \quad y'(0) = 15,$ $e^{-x} \cos x, e^{-x} \sin x$
20. **CAS PROJECT. Linear Independence.** Write a program for testing linear independence and dependence. Try it out on some of the problems in this and the next problem set and on examples of your own.

2.2 Homogeneous Linear ODEs with Constant Coefficients

We shall now consider second-order homogeneous linear ODEs whose coefficients a and b are constant,

$$y'' + ay' + by = 0. \tag{1}$$

These equations have important applications in mechanical and electrical vibrations, as we shall see in Secs. 2.4, 2.8, and 2.9.

To solve (1), we recall from Sec. 1.5 that the solution of the first-order linear ODE with a constant coefficient k

$$y' + ky = 0$$

is an exponential function $y = ce^{-kx}$. This gives us the idea to try as a solution of (1) the function

$$y = e^{\lambda x}. \tag{2}$$

Substituting (2) and its derivatives

$$y' = \lambda e^{\lambda x} \qquad \text{and} \qquad y'' = \lambda^2 e^{\lambda x}$$

into our equation (1), we obtain

$$(\lambda^2 + a\lambda + b)e^{\lambda x} = 0.$$

Hence if λ is a solution of the important **characteristic equation** (or *auxiliary equation*)

$$\lambda^2 + a\lambda + b = 0 \tag{3}$$

then the exponential function (2) is a solution of the ODE (1). Now from algebra we recall that the roots of this quadratic equation (3) are

$$\lambda_1 = \tfrac{1}{2}\left(-a + \sqrt{a^2 - 4b}\right), \qquad \lambda_2 = \tfrac{1}{2}\left(-a - \sqrt{a^2 - 4b}\right). \tag{4}$$

(3) and (4) will be basic because our derivation shows that the functions

$$y_1 = e^{\lambda_1 x} \qquad \text{and} \qquad y_2 = e^{\lambda_2 x} \tag{5}$$

are solutions of (1). Verify this by substituting (5) into (1).

From algebra we further know that the quadratic equation (3) may have three kinds of roots, depending on the sign of the discriminant $a^2 - 4b$, namely,

(Case I) *Two real roots if* $a^2 - 4b > 0$,
(Case II) *A real double root if* $a^2 - 4b = 0$,
(Case III) *Complex conjugate roots if* $a^2 - 4b < 0$.

Case I. Two Distinct Real-Roots λ_1 and λ_2

In this case, a basis of solutions of (1) on any interval is

$$y_1 = e^{\lambda_1 x} \qquad \text{and} \qquad y_2 = e^{\lambda_2 x}$$

because y_1 and y_2 are defined (and real) for all x and their quotient is not constant. The corresponding general solution is

$$y = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}. \tag{6}$$

EXAMPLE 1 General Solution in the Case of Distinct Real Roots

We can now solve $y'' - y = 0$ in Example 6 of Sec. 2.1 systematically. The characteristic equation is $\lambda^2 - 1 = 0$. Its roots are $\lambda_1 = 1$ and $\lambda_2 = -1$. Hence a basis of solutions is e^x and e^{-x} and gives the same general solution as before,

$$y = c_1 e^x + c_2 e^{-x}.$$

■

EXAMPLE 2 Initial Value Problem in the Case of Distinct Real Roots

Solve the initial value problem

$$y'' + y' - 2y = 0, \qquad y(0) = 4, \qquad y'(0) = -5.$$

Solution. ***Step 1. General solution.*** The characteristic equation is

$$\lambda^2 + \lambda - 2 = 0.$$

Its roots are

$$\lambda_1 = \tfrac{1}{2}(-1 + \sqrt{9}) = 1 \qquad \text{and} \qquad \lambda_2 = \tfrac{1}{2}(-1 - \sqrt{9}) = -2$$

so that we obtain the general solution

$$y = c_1 e^x + c_2 e^{-2x}.$$

Step 2. Particular solution. Since $y'(x) = c_1 e^x - 2c_2 e^{-2x}$, we obtain from the general solution and the initial conditions

$$y(0) = c_1 + c_2 = 4,$$
$$y'(0) = c_1 - 2c_2 = -5.$$

Hence $c_1 = 1$ and $c_2 = 3$. This gives the *answer* $y = e^x + 3e^{-2x}$. Figure 30 shows that the curve begins at $y = 4$ with a negative slope (-5, but note that the axes have different scales!), in agreement with the initial conditions. ■

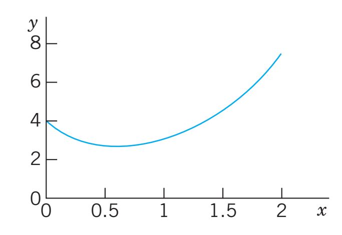


Fig. 30. Solution in Example 2

Case II. Real Double Root $\lambda = -a/2$

If the discriminant $a^2 - 4b$ is zero, we see directly from (4) that we get only one root, $\lambda = \lambda_1 = \lambda_2 = -a/2$, hence only one solution,

$$y_1 = e^{-(a/2)x}.$$

To obtain a second independent solution y_2 (needed for a basis), we use the method of reduction of order discussed in the last section, setting $y_2 = uy_1$. Substituting this and its derivatives $y_2' = u'y_1 + uy_1'$ and y_2'' into (1), we first have

$$(u''y_1 + 2u'y_1' + uy_1'') + a(u'y_1 + uy_1') + buy_1 = 0.$$

Collecting terms in u'', u', and u, as in the last section, we obtain

$$u''y_1 + u'(2y_1' + ay_1) + u(y_1'' + ay_1' + by_1) = 0.$$

The expression in the last parentheses is zero, since y_1 is a solution of (1). The expression in the first parentheses is zero, too, since

$$2y_1' = -ae^{-ax/2} = -ay_1.$$

We are thus left with $u''y_1 = 0$. Hence $u'' = 0$. By two integrations, $u = c_1x + c_2$. To get a second independent solution $y_2 = uy_1$, we can simply choose $c_1 = 1$, $c_2 = 0$ and take $u = x$. Then $y_2 = xy_1$. Since these solutions are not proportional, they form a basis. Hence in the case of a double root of (3) a basis of solutions of (1) on any interval is

$$e^{-ax/2}, \qquad xe^{-ax/2}.$$

The corresponding general solution is

$$y = (c_1 + c_2x)e^{-ax/2}. \tag{7}$$

WARNING! If λ is a *simple* root of (4), then $(c_1 + c_2x)e^{\lambda x}$ with $c_2 \neq 0$ is ***not*** a solution of (1).

EXAMPLE 3 General Solution in the Case of a Double Root

The characteristic equation of the ODE $y'' + 6y' + 9y = 0$ is $\lambda^2 + 6\lambda + 9 = (\lambda + 3)^2 = 0$. It has the double root $\lambda = -3$. Hence a basis is e^{-3x} and xe^{-3x}. The corresponding general solution is $y = (c_1 + c_2x)e^{-3x}$. ■

EXAMPLE 4 Initial Value Problem in the Case of a Double Root

Solve the initial value problem

$$y'' + y' + 0.25y = 0, \qquad y(0) = 3.0, \quad y'(0) = -3.5.$$

Solution. The characteristic equation is $\lambda^2 + \lambda + 0.25 = (\lambda + 0.5)^2 = 0$. It has the double root $\lambda = -0.5$. This gives the general solution

$$y = (c_1 + c_2x)e^{-0.5x}.$$

We need its derivative

$$y' = c_2e^{-0.5x} - 0.5(c_1 + c_2x)e^{-0.5x}.$$

From this and the initial conditions we obtain

$$y(0) = c_1 = 3.0, \qquad y'(0) = c_2 - 0.5c_1 = 3.5; \qquad \text{hence} \qquad c_2 = -2.$$

The particular solution of the initial value problem is $y = (3 - 2x)e^{-0.5x}$. See Fig. 31. ■

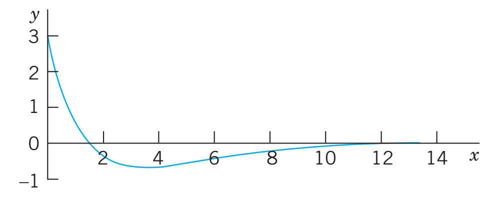


Fig. 31. Solution in Example 4

Case III. Complex Roots $-\frac{1}{2}a + i\omega$ and $-\frac{1}{2}a - i\omega$

This case occurs if the discriminant $a^2 - 4b$ of the characteristic equation (3) is negative. In this case, the roots of (3) are the complex $\lambda = -\frac{1}{2}a \pm i\omega$ that give the complex solutions of the ODE (1). However, we will show that we can obtain a basis of *real* solutions

$$y_1 = e^{-ax/2} \cos \omega x, \qquad y_2 = e^{-ax/2} \sin \omega x \qquad (\omega > 0) \tag{8}$$

where $\omega^2 = b - \frac{1}{4}a^2$. It can be verified by substitution that these are solutions in the present case. We shall derive them systematically after the two examples by using the complex exponential function. They form a basis on any interval since their quotient $\cot \omega x$ is not constant. Hence a real general solution in Case III is

$$y = e^{-ax/2}(A \cos \omega x + B \sin \omega x) \qquad (A, B \text{ arbitrary}). \tag{9}$$

EXAMPLE 5 Complex Roots. Initial Value Problem

Solve the initial value problem

$$y'' + 0.4y' + 9.04y = 0, \qquad y(0) = 0, \qquad y'(0) = 3.$$

Solution. ***Step 1. General solution.*** The characteristic equation is $\lambda^2 + 0.4\lambda + 9.04 = 0$. It has the roots $-0.2 \pm 3i$. Hence $\omega = 3$, and a general solution (9) is

$$y = e^{-0.2x}(A \cos 3x + B \sin 3x).$$

Step 2. Particular solution. The first initial condition gives $y(0) = A = 0$. The remaining expression is $y = Be^{-0.2x} \sin 3x$. We need the derivative (chain rule!)

$$y' = B(-0.2e^{-0.2x} \sin 3x + 3e^{-0.2x} \cos 3x).$$

From this and the second initial condition we obtain $y'(0) = 3B = 3$. Hence $B = 1$. Our solution is

$$y = e^{-0.2x} \sin 3x.$$

Figure 32 shows y and the curves of $e^{-0.2x}$ and $-e^{-0.2x}$ (dashed), between which the curve of y oscillates. Such "damped vibrations" (with $x = t$ being time) have important mechanical and electrical applications, as we shall soon see (in Sec. 2.4). ■

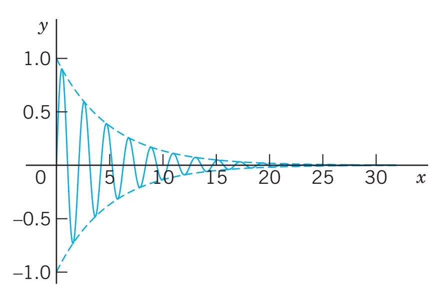


Fig. 32. Solution in Example 5

EXAMPLE 6 Complex Roots

A general solution of the ODE

$$y'' + \omega^2 y = 0 \qquad (\omega \text{ constant, not zero})$$

is

$$y = A \cos \omega x + B \sin \omega x.$$

With $\omega = 1$ this confirms Example 4 in Sec. 2.1. ■

Summary of Cases I–III

Case	Roots of (2)	Basis of (1)	General Solution of (1)
I	Distinct real λ_1, λ_2	$e^{\lambda_1 x}, e^{\lambda_2 x}$	$y = c_1 e^{\lambda_1 x} + c_2 e^{\lambda_2 x}$
II	Real double root $\lambda = -\frac{1}{2}a$	$e^{-ax/2}, xe^{-ax/2}$	$y = (c_1 + c_2 x)e^{-ax/2}$
III	Complex conjugate $\lambda_1 = -\frac{1}{2}a + i\omega$, $\lambda_2 = -\frac{1}{2}a - i\omega$	$e^{-ax/2}\cos \omega x$ $e^{-ax/2}\sin \omega x$	$y = e^{-ax/2}(A\cos \omega x + B \sin \omega x)$

It is very interesting that in applications to mechanical systems or electrical circuits, these three cases correspond to three different forms of motion or flows of current, respectively. We shall discuss this basic relation between theory and practice in detail in Sec. 2.4 (and again in Sec. 2.8).

Derivation in Case III. Complex Exponential Function

If verification of the solutions in (8) satisfies you, skip the systematic derivation of these real solutions from the complex solutions by means of the complex exponential function e^z of a complex variable $z = r + it$. We write $r + it$, not $x + iy$ because x and y occur in the ODE. The definition of e^z in terms of the real functions e^r, $\cos t$, and $\sin t$ is

$$e^z = e^{r+it} = e^r e^{it} = e^r(\cos t + i \sin t). \tag{10}$$

This is motivated as follows. For real $z = r$, hence $t = 0$, $\cos 0 = 1$, $\sin 0 = 0$, we get the real exponential function e^r. It can be shown that $e^{z_1+z_2} = e^{z_1}e^{z_2}$, just as in real. (Proof in Sec. 13.5.) Finally, if we use the Maclaurin series of e^z with $z = it$ as well as $i^2 = -1$, $i^3 = -i$, $i^4 = 1$, etc., and reorder the terms as shown (this is permissible, as can be proved), we obtain the series

$$\begin{aligned} e^{it} &= 1 + it + \frac{(it)^2}{2!} + \frac{(it)^3}{3!} + \frac{(it)^4}{4!} + \frac{(it)^5}{5!} + \cdots \\ &= 1 - \frac{t^2}{2!} + \frac{t^4}{4!} - + \cdots + i\left(t - \frac{t^3}{3!} + \frac{t^5}{5!} - + \cdots\right) \\ &= \cos t + i \sin t. \end{aligned}$$

(Look up these real series in your calculus book if necessary.) We see that we have obtained the formula

$$e^{it} = \cos t + i \sin t, \tag{11}$$

called the **Euler formula**. Multiplication by e^r gives (10).

For later use we note that $e^{-it} = \cos(-t) + i \sin(-t) = \cos t - i \sin t$, so that by addition and subtraction of this and (11),

$$\cos t = \tfrac{1}{2}(e^{it} + e^{-it}), \qquad \sin t = \frac{1}{2i}(e^{it} - e^{-it}). \tag{12}$$

After these comments on the definition (10), let us now turn to Case III.

In Case III the radicand $a^2 - 4b$ in (4) is negative. Hence $4b - a^2$ is positive and, using $\sqrt{-1} = i$, we obtain in (4)

$$\tfrac{1}{2}\sqrt{a^2 - 4b} = \tfrac{1}{2}\sqrt{-(4b - a^2)} = \sqrt{-(b - \tfrac{1}{4}a^2)} = i\sqrt{b - \tfrac{1}{4}a^2} = i\omega$$

with ω defined as in (8). Hence in (4),

$$\lambda_1 = \tfrac{1}{2}a + i\omega \qquad \text{and, similarly,} \qquad \lambda_2 = \tfrac{1}{2}a - i\omega.$$

Using (10) with $r = -\tfrac{1}{2}ax$ and $t = \omega x$, we thus obtain

$$e^{\lambda_1 x} = e^{-(a/2)x + i\omega x} = e^{-(a/2)x}(\cos \omega x + i \sin \omega x)$$
$$e^{\lambda_2 x} = e^{-(a/2)x - i\omega x} = e^{-(a/2)x}(\cos \omega x - i \sin \omega x).$$

We now add these two lines and multiply the result by $\tfrac{1}{2}$. This gives y_1 as in (8). Then we subtract the second line from the first and multiply the result by $1/(2i)$. This gives y_2 as in (8). These results obtained by addition and multiplication by constants are again solutions, as follows from the superposition principle in Sec. 2.1. This concludes the derivation of these real solutions in Case III.

PROBLEM SET 2.2

1–15 GENERAL SOLUTION

Find a general solution. Check your answer by substitution. ODEs of this kind have important applications to be discussed in Secs. 2.4, 2.7, and 2.9.

1. $4y'' - 25y = 0$
2. $y'' + 36y = 0$
3. $y'' + 6y' + 8.96y = 0$
4. $y'' + 4y' + (\pi^2 + 4)y = 0$
5. $y'' + 2\pi y' + \pi^2 y = 0$
6. $10y'' - 32y' + 25.6y = 0$
7. $y'' + 4.5y' = 0$
8. $y'' + y' + 3.25y = 0$
9. $y'' + 1.8y' - 2.08y = 0$
10. $100y'' + 240y' + (196\pi^2 + 144)y = 0$
11. $4y'' - 4y' - 3y = 0$
12. $y'' + 9y' + 20y = 0$
13. $9y'' - 30y' + 25y = 0$
14. $y'' + 2k^2 y' + k^4 y = 0$
15. $y'' + 0.54y' + (0.0729 + \pi)y = 0$

16–20 FIND AN ODE

$y'' + ay' + by = 0$ for the given basis.

16. $e^{2.6x}$, $e^{-4.3x}$
17. $e^{-\sqrt{5}x}$, $xe^{-\sqrt{5}x}$
18. $\cos 2\pi x$, $\sin 2\pi x$
19. $e^{(-2+i)x}$, $e^{(-2-i)x}$
20. $e^{-3.1x} \cos 2.1x$, $e^{-3.1x} \sin 2.1x$

21–30 INITIAL VALUES PROBLEMS

Solve the IVP. Check that your answer satisfies the ODE as well as the initial conditions. Show the details of your work.

21. $y'' + 25y = 0$, $y(0) = 4.6$, $y'(0) = -1.2$
22. The ODE in Prob. 4, $y(\tfrac{1}{2}) = 1$, $y'(\tfrac{1}{2}) = -2$
23. $y'' + y' - 6y = 0$, $y(0) = 10$, $y'(0) = 0$
24. $4y'' - 4y' - 3y = 0$, $y(-2) = e$, $y'(-2) = -e/2$
25. $y'' - y = 0$, $y(0) = 2$, $y'(0) = -2$
26. $y'' - k^2 y = 0$ $(k \neq 0)$, $y(0) = 1$, $y'(0) = 1$

27. The ODE in Prob. 5, $y(0) = 4.5, \quad y'(0) = -4.5\pi - 1 = 13.137$

28. $8y'' - 2y' - y = 0, \quad y(0) = -0.2, \quad y'(0) = -0.325$

29. The ODE in Prob. 15, $y(0) = 0, \quad y'(0) = 1$

30. $9y'' - 30y' + 25y = 0, \quad y(0) = 3.3, \quad y'(0) = 10.0$

31–36 LINEAR INDEPENDENCE is of basic importance, in this chapter, in connection with general solutions, as explained in the text. Are the following functions linearly independent on the given interval? Show the details of your work.

31. e^{kx}, xe^{kx}, any interval

32. $e^{ax}, e^{-ax}, \quad x > 0$

33. $x^2, x^2 \ln x, \quad x > 1$

34. $\ln x, \ln(x^3), \quad x > 1$

35. $\sin 2x, \cos x \sin x, \quad x < 0$

36. $e^{-x}\cos \frac{1}{2}x, 0, \quad -1 \leqq x \leqq 1$

37. Instability. Solve $y'' - y = 0$ for the initial conditions $y(0) = 1, y'(0) = -1$. Then change the initial conditions to $y(0) = 1.001, y'(0) = -0.999$ and explain why this small change of 0.001 at $t = 0$ causes a large change later, e.g., 22 at $t = 10$. This is instability: a small initial difference in setting a quantity (a current, for instance) becomes larger and larger with time t. This is undesirable.

38. TEAM PROJECT. General Properties of Solutions

(a) Coefficient formulas. Show how a and b in (1) can be expressed in terms of λ_1 and λ_2. Explain how these formulas can be used in constructing equations for given bases.

(b) Root zero. Solve $y'' + 4y' = 0$ (i) by the present method, and (ii) by reduction to first order. Can you explain why the result must be the same in both cases? Can you do the same for a general ODE $y'' + ay' = 0$?

(c) Double root. Verify directly that $xe^{\lambda x}$ with $\lambda = -a/2$ is a solution of (1) in the case of a double root. Verify and explain why $y = e^{-2x}$ is a solution of $y'' - y' - 6y = 0$ but xe^{-2x} is not.

(d) Limits. Double roots should be limiting cases of distinct roots λ_1, λ_2 as, say, $\lambda_2 \to \lambda_1$. Experiment with this idea. (Remember l'Hôpital's rule from calculus.) Can you arrive at $xe^{\lambda_1 x}$? Give it a try.

2.3 Differential Operators. *Optional*

This short section can be omitted without interrupting the flow of ideas. It will not be used subsequently, except for the notations Dy, D^2y, etc. to stand for y', y'', etc.

Operational calculus means the technique and application of operators. Here, an **operator** is a transformation that transforms a function into another function. Hence differential calculus involves an operator, the **differential operator** D, which transforms a (differentiable) function into its derivative. In operator notation we write $D = \frac{d}{dx}$ and

$$Dy = y' = \frac{dy}{dx}. \tag{1}$$

Similarly, for the higher derivatives we write $D^2y = D(Dy) = y''$, and so on. For example, $D \sin = \cos$, $D^2 \sin = -\sin$, etc.

For a homogeneous linear ODE $y'' + ay' + by = 0$ with constant coefficients we can now introduce the **second-order differential operator**

$$L = P(D) = D^2 + aD + bI,$$

where I is the **identity operator** defined by $Iy = y$. Then we can write that ODE as

$$Ly = P(D)y = (D^2 + aD + bI)y = 0. \tag{2}$$

P suggests "polynomial." L is a **linear operator**. By definition this means that if Ly and Lw exist (this is the case if y and w are twice differentiable), then $L(cy + kw)$ exists for any constants c and k, and

$$L(cy + kw) = cLy + kLw.$$

Let us show that from (2) we reach agreement with the results in Sec. 2.2. Since $(De^{\lambda})(x) = \lambda e^{\lambda x}$ and $(D^2 e^{\lambda})(x) = \lambda^2 e^{\lambda x}$, we obtain

$$(3) \qquad \begin{aligned} Le^{\lambda}(x) &= P(D)e^{\lambda}(x) = (D^2 + aD + bI)e^{\lambda}(x) \\ &= (\lambda^2 + a\lambda + b)e^{\lambda x} = P(\lambda)e^{\lambda x} = 0. \end{aligned}$$

This confirms our result of Sec. 2.2 that $e^{\lambda x}$ *is a solution of the ODE* (2) *if and only if* λ *is a solution of the characteristic equation* $P(\lambda) = 0$.

$P(\lambda)$ is a polynomial in the usual sense of algebra. If we replace λ by the operator D, we obtain the "operator polynomial" $P(D)$. *The point of this operational calculus is that* $P(D)$ *can be treated just like an algebraic quantity.* In particular, we can factor it.

EXAMPLE 1 **Factorization, Solution of an ODE**

Factor $P(D) = D^2 - 3D - 40I$ and solve $P(D)y = 0$.

Solution. $D^2 - 3D - 40I = (D - 8I)(D + 5I)$ because $I^2 = I$. Now $(D - 8I)y = y' - 8y = 0$ has the solution $y_1 = e^{8x}$. Similarly, the solution of $(D + 5I)y = 0$ is $y_2 = e^{-5x}$. This is a basis of $P(D)y = 0$ on any interval. From the factorization we obtain the ODE, as expected,

$$\begin{aligned} (D - 8I)(D + 5I)y &= (D - 8I)(y' + 5y) = D(y' + 5y) - 8(y' + 5y) \\ &= y'' + 5y' - 8y' - 40y = y'' - 3' - 40y = 0. \end{aligned}$$

Verify that this agrees with the result of our method in Sec. 2.2. This is not unexpected because we factored $P(D)$ in the same way as the characteristic polynomial $P(\lambda) = \lambda^2 - 3\lambda - 40$. ■

It was essential that L in (2) had *constant* coefficients. Extension of operator methods to variable-coefficient ODEs is more difficult and will not be considered here.

If operational methods were limited to the simple situations illustrated in this section, it would perhaps not be worth mentioning. Actually, the power of the operator approach appears in more complicated engineering problems, as we shall see in Chap. 6.

PROBLEM SET 2.3

1–5 APPLICATION OF DIFFERENTIAL OPERATORS

Apply the given operator to the given functions. Show all steps in detail.

1. $D^2 + 2D$; $\cosh 2x$, $e^{-x} + e^{2x}$, $\cos x$
2. $D - 3I$; $3x^2 + 3x$, $3e^{3x}$, $\cos 4x - \sin 4x$
3. $(D - 2I)^2$; e^{2x}, xe^{2x}, e^{-2x}
4. $(D + 6I)^2$; $6x + \sin 6x$, xe^{-6x}
5. $(D - 2I)(D + 3I)$; e^{2x}, xe^{2x}, e^{-3x}

6–12 GENERAL SOLUTION

Factor as in the text and solve.

6. $(D^2 + 4.00D + 3.36I)y = 0$
7. $(4D^2 - I)y = 0$
8. $(D^2 + 3I)y = 0$
9. $(D^2 - 4.20D + 4.41I)y = 0$
10. $(D^2 + 4.80D + 5.76I)y = 0$
11. $(D^2 - 4.00D + 3.84I)y = 0$
12. $(D^2 + 3.0D + 2.5I)y = 0$

13. Linear operator. Illustrate the linearity of L in (2) by taking $c = 4, k = -6, y = e^{2x}$, and $w = \cos 2x$. Prove that L is linear.

14. Double root. If $D^2 + aD + bI$ has distinct roots μ and λ, show that a particular solution is $y = (e^{\mu x} - e^{\lambda x})/(\mu - \lambda)$. Obtain from this a solution $xe^{\lambda x}$ by letting $\mu \to \lambda$ and applying l'Hôpital's rule.

15. Definition of linearity. Show that the definition of linearity in the text is equivalent to the following. If $L[y]$ and $L[w]$ exist, then $L[y + w]$ exists and $L[cy]$ and $L[kw]$ exist for all constants c and k, and $L[y + w] = L[y] + L[w]$ as well as $L[cy] = cL[y]$ and $L[kw] = kL[w]$.

2.4 Modeling of Free Oscillations of a Mass–Spring System

Linear ODEs with constant coefficients have important applications in mechanics, as we show in this section as well as in Sec. 2.8, and in electrical circuits as we show in Sec. 2.9. In this section we model and solve a basic mechanical system consisting of a mass on an elastic spring (a so-called "mass–spring system," Fig. 33), which moves up and down.

Setting Up the Model

We take an ordinary coil spring that resists extension as well as compression. We suspend it vertically from a fixed support and attach a body at its lower end, for instance, an iron ball, as shown in Fig. 33. We let $y = 0$ denote the position of the ball when the system is at rest (Fig. 33b). Furthermore, we choose ***the downward direction as positive***, thus regarding downward forces as *positive* and upward forces as *negative*.

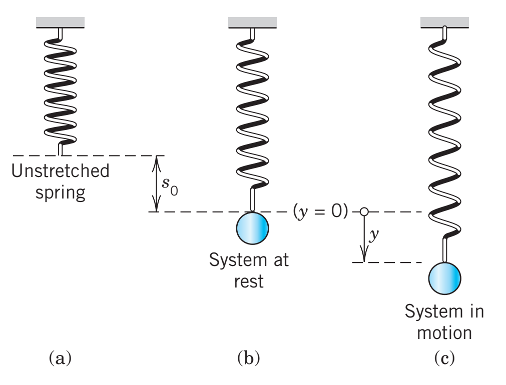


Fig. 33. Mechanical mass–spring system

We now let the ball move, as follows. We pull it down by an amount $y > 0$ (Fig. 33c). This causes a spring force

$$F_1 = -ky \quad \text{(1)} \qquad \textbf{(Hooke's law}^2\textbf{)}$$

proportional to the stretch y, with $k\,(>0)$ called the **spring constant**. The minus sign indicates that F_1 points upward, against the displacement. It is a *restoring force*: It wants to restore the system, that is, to pull it back to $y = 0$. Stiff springs have large k.

[2]ROBERT HOOKE (1635–1703), English physicist, a forerunner of Newton with respect to the law of gravitation.

Note that an additional force $-F_0$ is present in the spring, caused by stretching it in fastening the ball, but F_0 has no effect on the motion because it is in equilibrium with the weight W of the ball, $-F_0 = W = mg$, where $g = 980 \text{ cm/sec}^2 = 9.8 \text{ m/sec}^2 = 32.17 \text{ ft/sec}^2$ is the **constant of gravity at the Earth's surface** (not to be confused with the *universal gravitational constant* $G = gR^2/M = 6.67 \cdot 10^{-11} \text{ nt m}^2/\text{kg}^2$, which we shall not need; here $R = 6.37 \cdot 10^6$ m and $M = 5.98 \cdot 10^{24}$ kg are the Earth's radius and mass, respectively).

The motion of our mass–spring system is determined by **Newton's second law**

$$\text{Mass} \times \text{Acceleration} = my'' = \text{Force} \tag{2}$$

where $y'' = d^2y/dt^2$ and "Force" is the resultant of all the forces acting on the ball. (For systems of units, see the inside of the front cover.)

ODE of the Undamped System

Every system has damping. Otherwise it would keep moving forever. But if the damping is small and the motion of the system is considered over a relatively short time, we may disregard damping. Then Newton's law with $F = -F_1$ gives the model $my'' = -F_1 = -ky$; thus

$$my'' + ky = 0. \tag{3}$$

This is a homogeneous linear ODE with constant coefficients. A general solution is obtained as in Sec. 2.2, namely (see Example 6 in Sec. 2.2)

$$y(t) = A \cos \omega_0 t + B \sin \omega_0 t \qquad \omega_0 = \sqrt{\frac{k}{m}}. \tag{4}$$

This motion is called a **harmonic oscillation** (Fig. 34). Its *frequency* is $f = \omega_0/2\pi$ Hertz[3] (= cycles/sec) because cos and sin in (4) have the period $2\pi/\omega_0$. The frequency f is called the **natural frequency** of the system. (We write ω_0 to reserve ω for Sec. 2.8.)

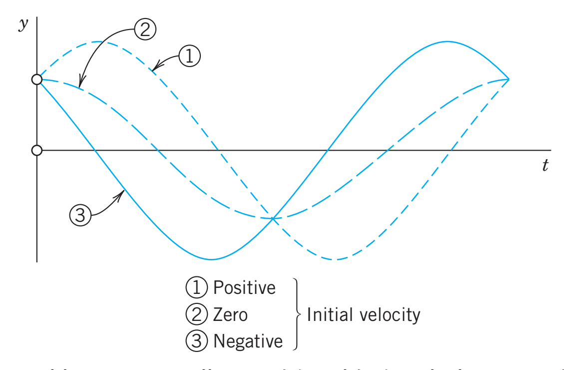


Fig. 34. Typical harmonic oscillations (4) and (4*) with the same $y(0) = A$ and different initial velocities $y'(0) = \omega_0 B$, positive ①, zero ②, negative ③

[3]**HEINRICH HERTZ (1857–1894)**, German physicist, who discovered electromagnetic waves, as the basis of wireless communication developed by GUGLIELMO MARCONI (1874–1937), Italian physicist (Nobel prize in 1909).

An alternative representation of (4), which shows the physical characteristics of amplitude and phase shift of (4), is

$$y(t) = C \cos(\omega_0 t - \delta) \tag{4*}$$

with $C = \sqrt{A^2 + B^2}$ and phase angle δ, where $\tan \delta = B/A$. This follows from the addition formula (6) in App. 3.1.

EXAMPLE 1 **Harmonic Oscillation of an Undamped Mass–Spring System**

If a mass–spring system with an iron ball of weight $W = 98$ nt (about 22 lb) can be regarded as undamped, and the spring is such that the ball stretches it 1.09 m (about 43 in.), how many cycles per minute will the system execute? What will its motion be if we pull the ball down from rest by 16 cm (about 6 in.) and let it start with zero initial velocity?

Solution. Hooke's law (1) with W as the force and 1.09 meter as the stretch gives $W = 1.09k$; thus $k = W/1.09 = 98/1.09 = 90$ [kg/sec^2] $= 90$ [nt/meter]. The mass is $m = W/g = 98/9.8 = 10$ [kg]. This gives the frequency $\omega_0/(2\pi) = \sqrt{k/m}/(2\pi) = 3/(2\pi) = 0.48$ [Hz] $= 29$ [cycles/min].

From (4) and the initial conditions, $y(0) = A = 0.16$ [meter] and $y'(0) = \omega_0 B = 0$. Hence the motion is

$$y(t) = 0.16 \cos 3t \text{ [meter]} \qquad \text{or} \qquad 0.52 \cos 3t \text{ [ft]} \qquad \text{(Fig. 35).}$$

If you have a chance of experimenting with a mass–spring system, don't miss it. You will be surprised about the good agreement between theory and experiment, usually within a fraction of one percent if you measure carefully. ■

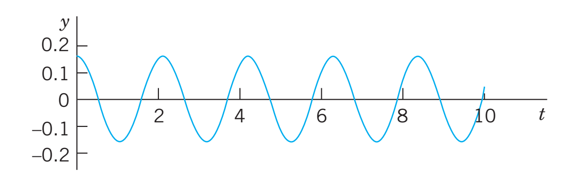


Fig. 35. Harmonic oscillation in Example 1

ODE of the Damped System

To our model $my'' = -ky$ we now add a damping force

$$F_2 = -cy',$$

obtaining $my'' = -ky - cy'$; thus the ODE of the damped mass–spring system is

$$my'' + cy' + ky = 0. \tag{5}$$ (Fig. 36)

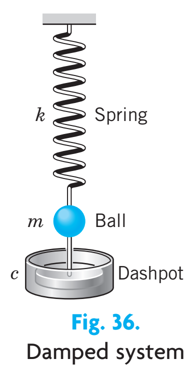


Fig. 36. Damped system

Physically this can be done by connecting the ball to a dashpot; see Fig. 36. We assume this damping force to be proportional to the velocity $y' = dy/dt$. This is generally a good approximation for small velocities.

The constant c is called the *damping constant.* Let us show that c is positive. Indeed, the damping force $F_2 = -cy'$ acts *against* the motion; hence for a downward motion we have $y' > 0$ which for positive c makes F negative (an upward force), as it should be. Similarly, for an upward motion we have $y' < 0$ which, for $c > 0$ makes F_2 positive (a downward force).

The ODE (5) is homogeneous linear and has constant coefficients. Hence we can solve it by the method in Sec. 2.2. The characteristic equation is (divide (5) by m)

$$\lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

By the usual formula for the roots of a quadratic equation we obtain, as in Sec. 2.2,

$$\text{(6)} \quad \lambda_1 = -\alpha + \beta, \quad \lambda_2 = -\alpha - \beta, \quad \text{where} \quad \alpha = \frac{c}{2m} \quad \text{and} \quad \beta = \frac{1}{2m}\sqrt{c^2 - 4mk}.$$

It is now interesting that depending on the amount of damping present—whether a lot of damping, a medium amount of damping or little damping—three types of motions occur, respectively:

Case I.	$c^2 > 4mk$.	*Distinct real roots* λ_1, λ_2.	**(Overdamping)**
Case II.	$c^2 = 4mk$.	*A real double root.*	**(Critical damping)**
Case III.	$c^2 < 4mk$.	*Complex conjugate roots.*	**(Underdamping)**

They correspond to the three Cases I, II, III in Sec. 2.2.

Discussion of the Three Cases

Case I. Overdamping

If the damping constant c is so large that $c^2 > 4mk$, then λ_1 and λ_2 are distinct real roots. In this case the corresponding general solution of (5) is

$$\textbf{(7)} \qquad y(t) = c_1 e^{-(\alpha-\beta)t} + c_2 e^{-(\alpha+\beta)t}.$$

We see that in this case, damping takes out energy so quickly that the body does not oscillate. For $t > 0$ both exponents in (7) are negative because $\alpha > 0, \beta > 0$, and $\beta^2 = \alpha^2 - k/m < \alpha^2$. Hence both terms in (7) approach zero as $t \to \infty$. Practically speaking, after a sufficiently long time the mass will be at rest at the *static equilibrium position* ($y = 0$). Figure 37 shows (7) for some typical initial conditions.

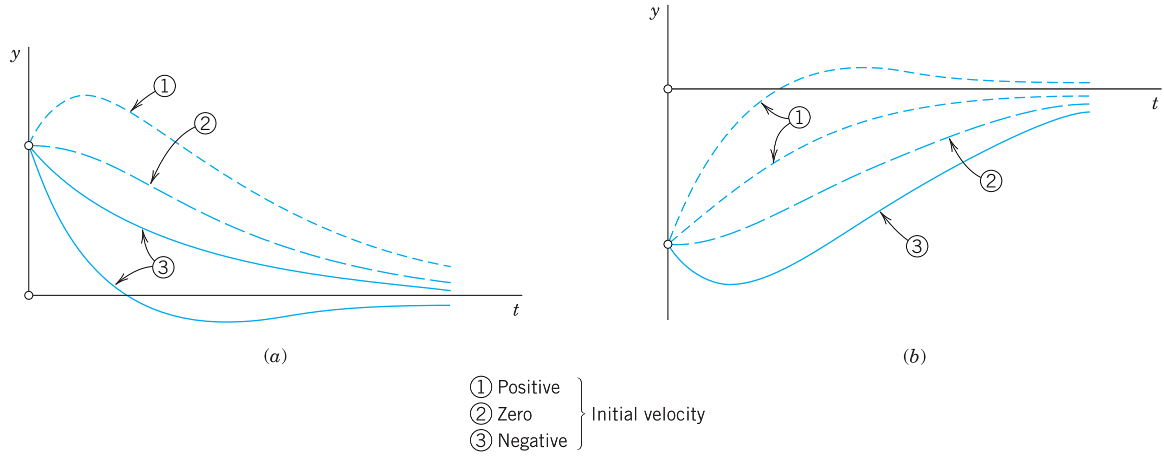


Fig. 37. Typical motions (7) in the overdamped case
(*a*) Positive initial displacement
(*b*) Negative initial displacement

Case II. Critical Damping

Critical damping is the border case between nonoscillatory motions (Case I) and oscillations (Case III). It occurs if the characteristic equation has a double root, that is, if $c^2 = 4mk$, so that $\beta = 0$, $\lambda_1 = \lambda_2 = -\alpha$. Then the corresponding general solution of (5) is

$$y(t) = (c_1 + c_2 t)e^{-\alpha t}. \tag{8}$$

This solution can pass through the equilibrium position $y = 0$ at most once because $e^{-\alpha t}$ is never zero and $c_1 + c_2 t$ can have at most one positive zero. If both c_1 and c_2 are positive (or both negative), it has no positive zero, so that y does not pass through 0 at all. Figure 38 shows typical forms of (8). Note that they look almost like those in the previous figure.

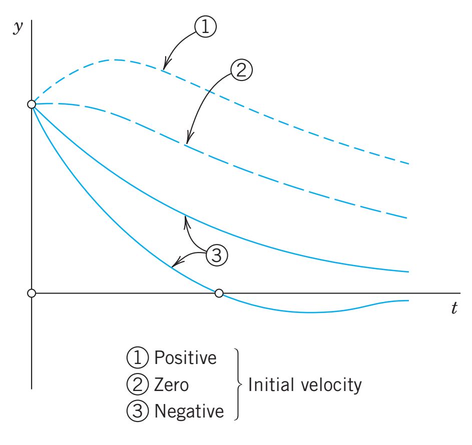


Fig. 38. Critical damping [see (8)]

Case III. Underdamping

This is the most interesting case. It occurs if the damping constant c is so small that $c^2 < 4mk$. Then β in (6) is no longer real but pure imaginary, say,

$$\text{(9)} \qquad \beta = i\omega^* \qquad \text{where} \qquad \omega^* = \frac{1}{2m}\sqrt{4mk - c^2} = \sqrt{\frac{k}{m} - \frac{c^2}{4m^2}} \qquad (>0).$$

(We now write ω^* to reserve ω for driving and electromotive forces in Secs. 2.8 and 2.9.) The roots of the characteristic equation are now complex conjugates,

$$\lambda_1 = -\alpha + i\omega^*, \qquad \lambda_2 = -\alpha - i\omega^*$$

with $\alpha = c/(2m)$, as given in (6). Hence the corresponding general solution is

$$\textbf{(10)} \qquad y(t) = e^{-\alpha t}(A \cos \omega^* t + B \sin \omega^* t) = Ce^{-\alpha t} \cos(\omega^* t - \delta)$$

where $C^2 = A^2 + B^2$ and $\tan \delta = B/A$, as in (4*).

This represents **damped oscillations**. Their curve lies between the dashed curves $y = Ce^{-\alpha t}$ and $y = -Ce^{-\alpha t}$ in Fig. 39, touching them when $\omega^* t - \delta$ is an integer multiple of π because these are the points at which $\cos(\omega^* t - \delta)$ equals 1 or -1.

The frequency is $\omega^*/(2\pi)$ Hz (hertz, cycles/sec). From (9) we see that the smaller c (>0) is, the larger is ω^* and the more rapid the oscillations become. If c approaches 0, then ω^* approaches $\omega_0 = \sqrt{k/m}$, giving the harmonic oscillation (4), whose frequency $\omega_0/(2\pi)$ is the natural frequency of the system.

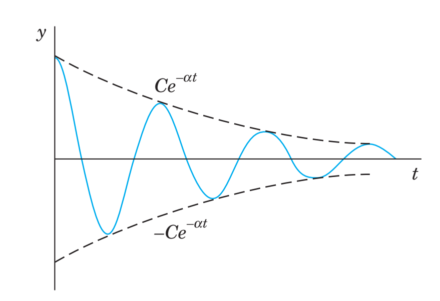


Fig. 39. Damped oscillation in Case III [see (10)]

EXAMPLE 2 The Three Cases of Damped Motion

How does the motion in Example 1 change if we change the damping constant c from one to another of the following three values, with $y(0) = 0.16$ and $y'(0) = 0$ as before?

$$\text{(I) } c = 100 \text{ kg/sec}, \qquad \text{(II) } c = 60 \text{ kg/sec}, \qquad \text{(III) } c = 10 \text{ kg/sec}.$$

Solution. It is interesting to see how the behavior of the system changes due to the effect of the damping, which takes energy from the system, so that the oscillations decrease in amplitude (Case III) or even disappear (Cases II and I).

(I) With $m = 10$ and $k = 90$, as in Example 1, the model is the initial value problem

$$10y'' + 100y' + 90y = 0, \qquad y(0) = 0.16 \text{ [meter]}, \qquad y'(0) = 0.$$

The characteristic equation is $10\lambda^2 + 100\lambda + 90 = 10(\lambda + 9)(\lambda + 1) = 0$. It has the roots -9 and -1. This gives the general solution

$$y = c_1 e^{-9t} + c_2 e^{-t}. \qquad \text{We also need} \qquad y' = -9c_1 e^{-9t} - c_2 e^{-t}.$$

The initial conditions give $c_1 + c_2 = 0.16$, $-9c_1 - c_2 = 0$. The solution is $c_1 = -0.02$, $c_2 = 0.18$. Hence in the overdamped case the solution is

$$y = -0.02e^{-9t} + 0.18e^{-t}.$$

It approaches 0 as $t \to \infty$. The approach is rapid; after a few seconds the solution is practically 0, that is, the iron ball is at rest.

(II) The model is as before, with $c = 60$ instead of 100. The characteristic equation now has the form $10\lambda^2 + 60\lambda + 90 = 10(\lambda + 3)^2 = 0$. It has the double root -3. Hence the corresponding general solution is

$$y = (c_1 + c_2 t)e^{-3t}. \qquad \text{We also need} \qquad y' = (c_2 - 3c_1 - 3c_2 t)e^{-3t}.$$

The initial conditions give $y(0) = c_1 = 0.16$, $y'(0) = c_2 - 3c_1 = 0$, $c_2 = 0.48$. Hence in the critical case the solution is

$$y = (0.16 + 0.48t)e^{-3t}.$$

It is always positive and decreases to 0 in a monotone fashion.

(III) The model now is $10y'' + 10y' + 90y = 0$. Since $c = 10$ is smaller than the critical c, we shall get oscillations. The characteristic equation is $10\lambda^2 + 10\lambda + 90 = 10[(\lambda + \frac{1}{2})^2 + 9 - \frac{1}{4}] = 0$. It has the complex roots [see (4) in Sec. 2.2 with $a = 1$ and $b = 9$]

$$\lambda = -0.5 \pm \sqrt{0.5^2 - 9} = -0.5 \pm 2.96i.$$

This gives the general solution

$$y = e^{-0.5t}(A \cos 2.96t + B \sin 2.96t).$$

Thus $y(0) = A = 0.16$. We also need the derivative

$$y' = e^{-0.5t}(-0.5A \cos 2.96t - 0.5B \sin 2.96t - 2.96A \sin 2.96t + 2.96B \cos 2.96t).$$

Hence $y'(0) = -0.5A + 2.96B = 0$, $\quad B = 0.5A/2.96 = 0.027$. This gives the solution

$$y = e^{-0.5t}(0.16 \cos 2.96t + 0.027 \sin 2.96t) = 0.162e^{-0.5t} \cos (2.96t - 0.17).$$

We see that these damped oscillations have a smaller frequency than the harmonic oscillations in Example 1 by about 1% (since 2.96 is smaller than 3.00 by about 1%). Their amplitude goes to zero. See Fig. 40. ■

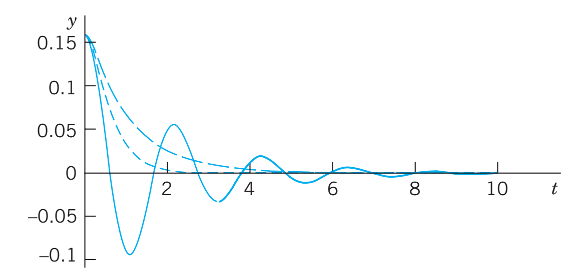


Fig. 40. The three solutions in Example 2

This section concerned ***free motions*** of mass–spring systems. Their models are *homogeneous* linear ODEs. *Nonhomogeneous* linear ODEs will arise as models of **forced motions**, that is, motions under the influence of a “driving force.” We shall study them in Sec. 2.8, after we have learned how to solve those ODEs.

PROBLEM SET 2.4

1–10 HARMONIC OSCILLATIONS (UNDAMPED MOTION)

1. **Initial value problem.** Find the harmonic motion (4) that starts from y_0 with initial velocity v_0. Graph or sketch the solutions for $\omega_0 = \pi$, $y_0 = 1$, and various v_0 of your choice on common axes. At what t-values do all these curves intersect? Why?
2. **Frequency.** If a weight of 20 nt (about 4.5 lb) stretches a certain spring by 2 cm, what will the frequency of the corresponding harmonic oscillation be? The period?
3. **Frequency.** How does the frequency of the harmonic oscillation change if we (i) double the mass, (ii) take a spring of twice the modulus? First find qualitative answers by physics, then look at formulas.
4. **Initial velocity.** Could you make a harmonic oscillation move faster by giving the body a greater initial push?
5. **Springs in parallel.** What are the frequencies of vibration of a body of mass $m = 5$ kg (i) on a spring of modulus $k_1 = 20$ nt/m, (ii) on a spring of modulus $k_2 = 45$ nt/m, (iii) on the two springs in parallel? See Fig. 41.

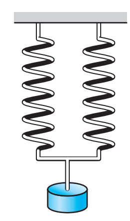


Fig. 41. Parallel springs (Problem 5)

6. **Spring in series.** If a body hangs on a spring s_1 of modulus $k_1 = 8$, which in turn hangs on a spring s_2 of modulus $k_2 = 12$, what is the modulus k of this combination of springs?
7. **Pendulum.** Find the frequency of oscillation of a pendulum of length L (Fig. 42), neglecting air resistance and the weight of the rod, and assuming θ to be so small that $\sin\theta$ practically equals θ.

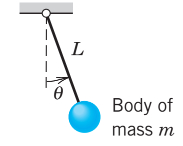


Fig. 42. Pendulum (Problem 7)

8. **Archimedian principle.** This principle states that the buoyancy force equals the weight of the water displaced by the body (partly or totally submerged). The cylindrical buoy of diameter 60 cm in Fig. 43 is floating in water with its axis vertical. When depressed downward in the water and released, it vibrates with period 2 sec. What is its weight?

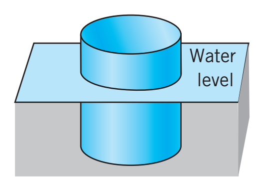


Fig. 43. Buoy (Problem 8)

9. **Vibration of water in a tube.** If 1 liter of water (about 1.06 US quart) is vibrating up and down under the influence of gravitation in a U-shaped tube of diameter 2 cm (Fig. 44), what is the frequency? Neglect friction. First guess.

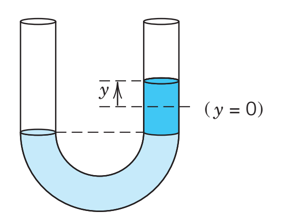


Fig. 44. Tube (Problem 9)

10. **TEAM PROJECT. Harmonic Motions of Similar Models.** The ***unifying power of mathematical methods*** results to a large extent from the fact that different physical (or other) systems may have the same or very similar models. Illustrate this for the following three systems

    (a) Pendulum clock. A clock has a 1-meter pendulum. The clock ticks once for each time the pendulum completes a full swing, returning to its original position. How many times a minute does the clock tick?

    (b) Flat spring (Fig. 45). The harmonic oscillations of a flat spring with a body attached at one end and horizontally clamped at the other are also governed by (3). Find its motions, assuming that the body weighs 8 nt (about 1.8 lb), the system has its static equilibrium 1 cm below the horizontal line, and we let it start from this position with initial velocity 10 cm/sec.

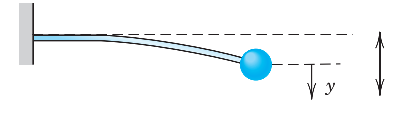


Fig. 45. Flat spring

(c) Torsional vibrations (Fig. 46). Undamped torsional vibrations (rotations back and forth) of a wheel attached to an elastic thin rod or wire are governed by the equation $I_0\theta'' + K\theta = 0$, where θ is the angle measured from the state of equilibrium. Solve this equation for $K/I_0 = 13.69 \text{ sec}^{-2}$, initial angle $30°(= 0.5235 \text{ rad})$ and initial angular velocity $20° \text{ sec}^{-1} (= 0.349 \text{ rad} \cdot \text{sec}^{-1})$.

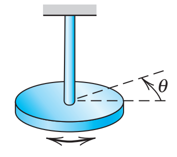


Fig. 46. Torsional vibrations

11–20 DAMPED MOTION

11. Overdamping. Show that for (7) to satisfy initial conditions $y(0) = y_0$ and $v(0) = v_0$ we must have $c_1 = [(1 + \alpha/\beta)y_0 + v_0/\beta]/2$ and $c_2 = [(1 - \alpha/\beta)y_0 - v_0/\beta]/2$.

12. Overdamping. Show that in the overdamped case, the body can pass through $y = 0$ at most once (Fig. 37).

13. Initial value problem. Find the critical motion (8) that starts from y_0 with initial velocity v_0. Graph solution curves for $\alpha = 1$, $y_0 = 1$ and several v_0 such that (i) the curve does not intersect the t-axis, (ii) it intersects it at $t = 1, 2, \dots, 5$, respectively.

14. Shock absorber. What is the smallest value of the damping constant of a shock absorber in the suspension of a wheel of a car (consisting of a spring and an absorber) that will provide (theoretically) an oscillation-free ride if the mass of the car is 2000 kg and the spring constant equals 4500 kg/sec^2?

15. Frequency. Find an approximation formula for ω^* in terms of ω_0 by applying the binomial theorem in (9) and retaining only the first two terms. How good is the approximation in Example 2, III?

16. Maxima. Show that the maxima of an underdamped motion occur at equidistant t-values and find the distance.

17. Underdamping. Determine the values of t corresponding to the maxima and minima of the oscillation $y(t) = e^{-t} \sin t$. Check your result by graphing $y(t)$.

18. Logarithmic decrement. Show that the ratio of two consecutive maximum amplitudes of a damped oscillation (10) is constant, and the natural logarithm of this ratio called the *logarithmic decrement*, equals $\Delta = 2\pi\alpha/\omega^*$. Find Δ for the solutions of $y'' + 2y' + 5y = 0$.

19. Damping constant. Consider an underdamped motion of a body of mass $m = 0.5$ kg. If the time between two consecutive maxima is 3 sec and the maximum amplitude decreases to $\frac{1}{2}$ its initial value after 10 cycles, what is the damping constant of the system?

20. CAS PROJECT. Transition Between Cases I, II, III. Study this transition in terms of graphs of typical solutions. (Cf. Fig. 47.)

(a) ***Avoiding unnecessary generality is part of good modeling.*** Show that the initial value problems (A) and (B),

$$\text{(A)} \quad y'' + cy' + y = 0, \quad y(0) = 1, \quad y'(0) = 0$$

(B) the same with different c and $y'(0) = -2$ (instead of 0), will give practically as much information as a problem with other m, k, $y(0)$, $y'(0)$.

(b) ***Consider* (A).** Choose suitable values of c, perhaps better ones than in Fig. 47, for the transition from Case III to II and I. Guess c for the curves in the figure.

(c) ***Time to go to rest.*** Theoretically, this time is infinite (why?). Practically, the system is at rest when its motion has become very small, say, less than 0.1% of the initial displacement (this choice being up to us), that is in our case,

$$(11) \quad |y(t)| < 0.001 \quad \text{for all } t \text{ greater than some } t_1.$$

In engineering constructions, damping can often be varied without too much trouble. Experimenting with your graphs, find empirically a relation between t_1 and c.

(d) ***Solve* (A) *analytically.*** Give a reason why the solution c of $y(t_2) = -0.001$, with t_2 the solution of $y'(t) = 0$, will give you the best possible c satisfying (11).

(e) Consider (B) empirically as in (a) and (b). What is the main difference between (B) and (A)?

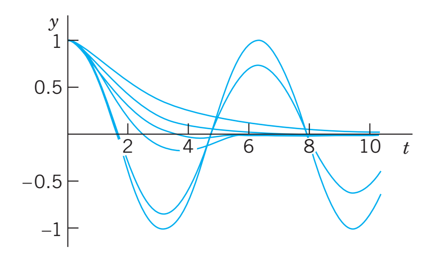


Fig. 47. CAS Project 20

2.5 Euler–Cauchy Equations

Euler–Cauchy equations[4] are ODEs of the form

$$x^2y'' + axy' + by = 0 \tag{1}$$

with given constants a and b and unknown function $y(x)$. We substitute

$$y = x^m, \qquad y' = mx^{m-1}, \qquad y'' = m(m-1)x^{m-2}$$

into (1). This gives

$$x^2m(m-1)x^{m-2} + axmx^{m-1} + bx^m = 0$$

and we now see that $y = x^m$ was a rather natural choice because we have obtained a common factor x^m. Dropping it, we have the auxiliary equation $m(m-1) + am + b = 0$ or

$$m^2 + (a-1)m + b = 0. \qquad \text{(Note: } a-1, \text{ not } a.) \tag{2}$$

Hence $y = x^m$ is a solution of (1) if and only if m is a root of (2). The roots of (2) are

$$m_1 = \tfrac{1}{2}(1-a) + \sqrt{\tfrac{1}{4}(1-a)^2 - b}, \qquad m_2 = \tfrac{1}{2}(1-a) - \sqrt{\tfrac{1}{4}(1-a)^2 - b}. \tag{3}$$

Case I. Real different roots m_1 and m_2 give two real solutions

$$y_1(x) = x^{m_1} \qquad \text{and} \qquad y_2(x) = x^{m_2}.$$

These are linearly independent since their quotient is not constant. Hence they constitute a basis of solutions of (1) for all x for which they are real. The corresponding general solution for all these x is

$$y = c_1x^{m_1} + c_2x^{m_2} \qquad (c_1, c_2 \text{ arbitrary}). \tag{4}$$

EXAMPLE 1 General Solution in the Case of Different Real Roots

The Euler–Cauchy equation $x^2y'' + 1.5xy' - 0.5y = 0$ has the auxiliary equation $m^2 + 0.5m - 0.5 = 0$. The roots are 0.5 and -1. Hence a basis of solutions for all positive x is $y_1 = x^{0.5}$ and $y_2 = 1/x$ and gives the general solution

$$y = c_1\sqrt{x} + \frac{c_2}{x} \qquad (x > 0). \quad \blacksquare$$

---

[4]**LEONHARD EULER (1707–1783)** was an enormously creative Swiss mathematician. He made fundamental contributions to almost all branches of mathematics and its application to physics. His important books on algebra and calculus contain numerous basic results of his own research. The great French mathematician AUGUSTIN LOUIS CAUCHY (1789–1857) is the father of modern analysis. He is the creator of complex analysis and had great influence on ODEs, PDEs, infinite series, elasticity theory, and optics.

Case II. A real double root $m_1 = \frac{1}{2}(1 - a)$ occurs if and only if $b = \frac{1}{4}(a - 1)^2$ because then (2) becomes $[m + \frac{1}{2}(a - 1)]^2$, as can be readily verified. Then a solution is $y_1 = x^{(1-a)/2}$, and (1) is of the form

$$x^2y'' + axy' + \tfrac{1}{4}(1 - a)^2 y = 0 \qquad \text{or} \qquad y'' + \frac{a}{x}y' + \frac{(1-a)^2}{4x^2}y = 0. \tag{5}$$

A second linearly independent solution can be obtained by the method of reduction of order from Sec. 2.1, as follows. Starting from $y_2 = uy_1$, we obtain for u the expression (9) Sec. 2.1, namely,

$$u = \int U\,dx \qquad \text{where} \qquad U = \frac{1}{y_1^2}\exp\left(-\int p\,dx\right).$$

From (5) in standard form (second ODE) we see that $p = a/x$ (not ax; this is essential!). Hence $\exp \int(-p\,dx) = \exp(-a \ln x) = \exp(\ln x^{-a}) = 1/x^a$. Division by $y_1^2 = x^{1-a}$ gives $U = 1/x$, so that $u = \ln x$ by integration. Thus, $y_2 = uy_1 = y_1 \ln x$, and y_1 and y_2 are linearly independent since their quotient is not constant. The general solution corresponding to this basis is

$$y = (c_1 + c_2 \ln x)\, x^m, \qquad m = \tfrac{1}{2}(1 - a). \tag{6}$$

EXAMPLE 2 General Solution in the Case of a Double Root

The Euler–Cauchy equation $x^2y'' - 5xy' + 9y = 0$ has the auxiliary equation $m^2 - 6m + 9 = 0$. It has the double root $m = 3$, so that a general solution for all positive x is

$$y = (c_1 + c_2 \ln x)\, x^3.$$ ■

Case III. Complex conjugate roots are of minor practical importance, and we discuss the derivation of real solutions from complex ones just in terms of a typical example.

EXAMPLE 3 Real General Solution in the Case of Complex Roots

The Euler–Cauchy equation $x^2y'' + 0.6xy' + 16.04y = 0$ has the auxiliary equation $m^2 - 0.4m + 16.04 = 0$. The roots are complex conjugate, $m_1 = 0.2 + 4i$ and $m_2 = 0.2 - 4i$, where $i = \sqrt{-1}$. We now use the trick of writing $x = e^{\ln x}$ and obtain

$$x^{m_1} = x^{0.2+4i} = x^{0.2}(e^{\ln x})^{4i} = x^{0.2}e^{(4 \ln x)i},$$
$$x^{m_2} = x^{0.2-4i} = x^{0.2}(e^{\ln x})^{-4i} = x^{0.2}e^{-(4 \ln x)i}.$$

Next we apply Euler's formula (11) in Sec. 2.2 with $t = 4 \ln x$ to these two formulas. This gives

$$x^{m_1} = x^{0.2}[\cos(4 \ln x) + i \sin(4 \ln x)],$$
$$x^{m_2} = x^{0.2}[\cos(4 \ln x) - i \sin(4 \ln x)].$$

We now add these two formulas, so that the sine drops out, and divide the result by 2. Then we subtract the second formula from the first, so that the cosine drops out, and divide the result by $2i$. This yields

$$x^{0.2}\cos(4 \ln x) \qquad \text{and} \qquad x^{0.2}\sin(4 \ln x)$$

respectively. By the superposition principle in Sec. 2.2 these are solutions of the Euler–Cauchy equation (1). Since their quotient $\cot(4 \ln x)$ is not constant, they are linearly independent. Hence they form a basis of solutions, and the corresponding real general solution for all positive x is

$$y = x^{0.2}[A \cos(4 \ln x) + B \sin(4 \ln x)]. \tag{8}$$

Figure 48 shows typical solution curves in the three cases discussed, in particular the real basis functions in Examples 1 and 3. ■

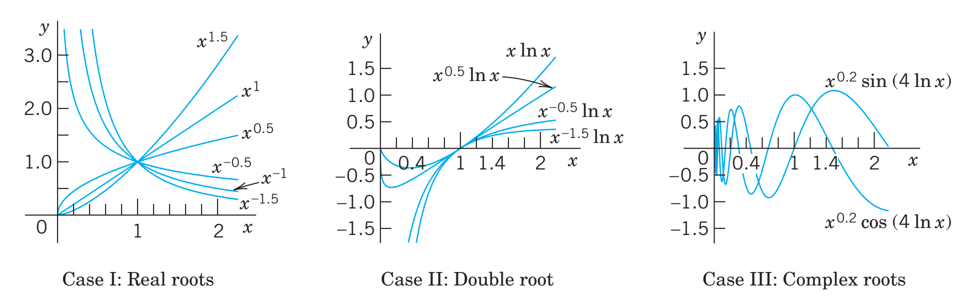


Fig. 48. Euler–Cauchy equations

EXAMPLE 4 Boundary Value Problem. Electric Potential Field Between Two Concentric Spheres

Find the electrostatic potential $v = v(r)$ between two concentric spheres of radii $r_1 = 5$ cm and $r_2 = 10$ cm kept at potentials $v_1 = 110$ V and $v_2 = 0$, respectively.

Physical Information. $v(r)$ is a solution of the Euler–Cauchy equation $rv'' + 2v' = 0$, where $v' = dv/dr$.

Solution. The auxiliary equation is $m^2 + m = 0$. It has the roots 0 and -1. This gives the general solution $v(r) = c_1 + c_2/r$. From the "boundary conditions" (the potentials on the spheres) we obtain

$$v(5) = c_1 + \frac{c_2}{5} = 110. \qquad v(10) = c_1 + \frac{c_2}{10} = 0.$$

By subtraction, $c_2/10 = 110$, $c_2 = 1100$. From the second equation, $c_1 = -c_2/10 = -110$. *Answer:* $v(r) = -110 + 1100/r$ V. Figure 49 shows that the potential is not a straight line, as it would be for a potential between two parallel plates. For example, on the sphere of radius 7.5 cm it is not $110/2 = 55$ V, but considerably less. (What is it?) ■

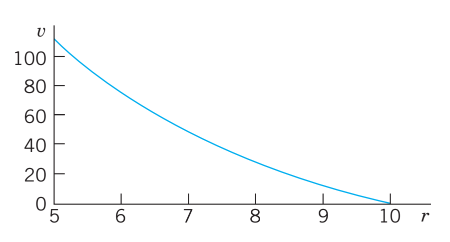


Fig. 49. Potential $v(r)$ in Example 4

PROBLEM SET 2.5

1. **Double root.** Verify directly by substitution that $x^{(1-a)/2} \ln x$ is a solution of (1) if (2) has a double root, but $x^{m_1} \ln x$ and $x^{m_2} \ln x$ are not solutions of (1) if the roots m_1 and m_2 of (2) are different.

2–11 GENERAL SOLUTION

Find a real general solution. Show the details of your work.

2. $x^2y'' - 20y = 0$
3. $5x^2y'' + 23xy' + 16.2y = 0$
4. $xy'' + 2y' = 0$
5. $4x^2y'' + 5y = 0$
6. $x^2y'' + 0.7xy' - 0.1y = 0$
7. $(x^2D^2 - 4xD + 6I)y = C$
8. $(x^2D^2 - 3xD + 4I)y = 0$
9. $(x^2D^2 - 0.2xD + 0.36I)y = 0$
10. $(x^2D^2 - xD + 5I)y = 0$
11. $(x^2D^2 - 3xD + 10I)y = 0$

12–19 **INITIAL VALUE PROBLEM**

Solve and graph the solution. Show the details of your work.

12. $x^2y'' - 4xy' + 6y = 0, \quad y(1) = 0.4, \quad y'(1) = 0$

13. $x^2y'' + 3xy' + 0.75y = 0, \quad y(1) = 1,$
$y'(1) = -1.5$

14. $x^2y'' + xy' + 9y = 0, \quad y(1) = 0, \quad y'(1) = 2.5$

15. $x^2y'' + 3xy' + y = 0, \quad y(1) = 3.6, \quad y'(1) = 0.4$

16. $(x^2D^2 - 3xD + 4I)y = 0, \quad y(1) = -\pi, \quad y'(1) = 2\pi$

17. $(x^2D^2 + xD + I)y = 0, \quad y(1) = 1, \quad y'(1) = 1$

18. $(9x^2D^2 + 3xD + I)y = 0, \quad y(1) = 1, \quad y'(1) = 0$

19. $(x^2D^2 - xD - 15I)y = 0, \quad y(1) = 0.1,$
$y'(1) = -4.5$

20. TEAM PROJECT. Double Root

(a) Derive a second linearly independent solution of (1) by reduction of order; but instead of using (9), Sec. 2.1, perform all steps directly for the present ODE (1).

(b) Obtain $x^m \ln x$ by considering the solutions x^m and x^{m+s} of a suitable Euler–Cauchy equation and letting $s \to 0$.

(c) Verify by substitution that $x^m \ln x$, $m = (1 - a)/2$, is a solution in the critical case.

(d) Transform the Euler–Cauchy equation (1) into an ODE with constant coefficients by setting $x = e^t$ $(x > 0)$.

(e) Obtain a second linearly independent solution of the Euler–Cauchy equation in the "critical case" from that of a constant-coefficient ODE.

2.6 Existence and Uniqueness of Solutions. Wronskian

In this section we shall discuss the general theory of homogeneous linear ODEs

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

with continuous, but otherwise arbitrary, ***variable coefficients*** p and q. This will concern the existence and form of a general solution of (1) as well as the uniqueness of the solution of initial value problems consisting of such an ODE and two initial conditions

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \tag{2}$$

with given x_0, K_0, and K_1.

The two main results will be Theorem 1, stating that such an initial value problem always has a solution which is unique, and Theorem 4, stating that a general solution

$$y = c_1y_1 + c_2y_2 \qquad (c_1, c_2 \text{ arbitrary}) \tag{3}$$

includes all solutions. Hence *linear* ODEs with continuous coefficients have no *"singular solutions"* (solutions not obtainable from a general solution).

Clearly, no such theory was needed for constant-coefficient or Euler–Cauchy equations because everything resulted explicitly from our calculations.

Central to our present discussion is the following theorem.

THEOREM 1

Existence and Uniqueness Theorem for Initial Value Problems

If $p(x)$ and $q(x)$ are continuous functions on some open interval I (see Sec. 1.1) *and x_0 is in I, then the initial value problem consisting of* (1) *and* (2) *has a unique solution $y(x)$ on the interval I.*

The proof of existence uses the same prerequisites as the existence proof in Sec. 1.7 and will not be presented here; it can be found in Ref. [A11] listed in App. 1. Uniqueness proofs are usually simpler than existence proofs. But for Theorem 1, even the uniqueness proof is long, and we give it as an additional proof in App. 4.

Linear Independence of Solutions

Remember from Sec. 2.1 that a general solution on an open interval I is made up from a **basis** y_1, y_2 on I, that is, from a pair of linearly independent solutions on I. Here we call y_1, y_2 **linearly independent** on I if the equation

$$k_1 y_1(x) + k_2 y_2(x) = 0 \quad \text{on } I \qquad \text{implies} \qquad k_1 = 0, \quad k_2 = 0. \tag{4}$$

We call y_1, y_2 **linearly dependent** on I if this equation also holds for constants k_1, k_2 not both 0. In this case, and only in this case, y_1 and y_2 are proportional on I, that is (see Sec. 2.1),

$$\text{(a)} \quad y_1 = k y_2 \qquad \text{or} \qquad \text{(b)} \quad y_2 = l y_1 \qquad \text{for all on } I. \tag{5}$$

For our discussion the following criterion of linear independence and dependence of solutions will be helpful.

THEOREM 2

Linear Dependence and Independence of Solutions

Let the ODE (1) *have continuous coefficients $p(x)$ and $q(x)$ on an open interval I. Then two solutions y_1 and y_2 of* (1) *on I are linearly dependent on I if and only if their* "**Wronskian**"

$$W(y_1, y_2) = y_1 y_2' - y_2 y_1' \tag{6}$$

is 0 *at some x_0 in I. Furthermore, if $W = 0$ at an $x = x_0$ in I, then $W = 0$ on I; hence, if there is an x_1 in I at which W is not* 0*, then y_1, y_2 are linearly independent on I.*

PROOF **(a)** Let y_1 and y_2 be linearly dependent on I. Then (5a) or (5b) holds on I. If (5a) holds, then

$$W(y_1, y_2) = y_1 y_2' - y_2 y_1' = k y_2 y_2' - y_2 k y_2' = 0.$$

Similarly if (5b) holds.

(b) Conversely, we let $W(y_1, y_2) = 0$ for some $x = x_0$ and show that this implies linear dependence of y_1 and y_2 on I. We consider the linear system of equations in the unknowns k_1, k_2

$$\begin{aligned} k_1 y_1(x_0) + k_2 y_2(x_0) &= 0 \\ k_1 y_1'(x_0) + k_2 y_2'(x_0) &= 0. \end{aligned} \tag{7}$$

To eliminate k_2, multiply the first equation by y_2' and the second by $-y_2$ and add the resulting equations. This gives

$$k_1y_1(x_0)y_2'(x_0) - k_1y_1'(x_0)y_2(x_0) = k_1W(y_1(x_0), y_2(x_0)) = 0.$$

Similarly, to eliminate k_1, multiply the first equation by $-y_1'$ and the second by y_1 and add the resulting equations. This gives

$$k_2W(y_1(x_0), y_2(x_0)) = 0.$$

If W were not 0 at x_0, we could divide by W and conclude that $k_1 = k_2 = 0$. Since W is 0, division is not possible, and the system has a solution for which k_1 and k_2 are not both 0. Using *these numbers* k_1, k_2, we introduce the function

$$y(x) = k_1y_1(x) + k_2y_2(x).$$

Since (1) is homogeneous linear, Fundamental Theorem 1 in Sec. 2.1 (the superposition principle) implies that this function is a solution of (1) on I. From (7) we see that it satisfies the initial conditions $y(x_0) = 0$, $y'(x_0) = 0$. Now another solution of (1) satisfying the same initial conditions is $y^* \equiv 0$. Since the coefficients p and q of (1) are continuous, Theorem 1 applies and gives uniqueness, that is, $y \equiv y^*$, written out

$$k_1y_1 + k_2y_2 \equiv 0 \qquad \text{on } I.$$

Now since k_1 and k_2 are not both zero, this means linear dependence of y_1, y_2 on I.

(c) We prove the last statement of the theorem. If $W(x_0) = 0$ at an x_0 in I, we have linear dependence of y_1, y_2 on I by part (b), hence $W \equiv 0$ by part (a) of this proof. Hence in the case of linear dependence it cannot happen that $W(x_1) \neq 0$ at an x_1 in I. If it does happen, it thus implies linear independence as claimed. ■

For calculations, the following formulas are often simpler than (6).

(6*) $$W(y_1, y_2) = (a) \quad \left(\frac{y_2}{y_1}\right)' y_1^2 \quad (y_1 \neq 0) \qquad \text{or} \qquad (b) \quad -\left(\frac{y_1}{y_2}\right)' y_2^2 \quad (y_2 \neq 0).$$

These formulas follow from the quotient rule of differentiation.

Remark. Determinants. Students familiar with second-order determinants may have noticed that

$$W(y_1, y_2) = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix} = y_1y_2' - y_2y_1'.$$

This determinant is called the *Wronski determinant*[5] or, briefly, the **Wronskian**, of two solutions y_1 and y_2 of (1), as has already been mentioned in (6). Note that its four entries occupy the same positions as in the linear system (7).

[5]Introduced by WRONSKI (JOSEF MARIA HÖNE, 1776–1853), Polish mathematician.

EXAMPLE 1 **Illustration of Theorem 2**

The functions $y_1 = \cos \omega x$ and $y_2 = \sin \omega x$ are solutions of $y'' + \omega^2 y = 0$. Their Wronskian is

$$W(\cos \omega x, \sin \omega x) = \begin{vmatrix} \cos \omega x & \sin \omega x \\ -\omega \sin \omega x & \omega \cos \omega x \end{vmatrix} = y_1 y_2' - y_2 y_1' = \omega \cos^2 \omega x + \omega \sin^2 \omega x = \omega.$$

Theorem 2 shows that these solutions are linearly independent if and only if $\omega \neq 0$. Of course, we can see this directly from the quotient $y_2/y_1 = \tan \omega x$. For $\omega = 0$ we have $y_2 = 0$, which implies linear dependence (why?). ■

EXAMPLE 2 **Illustration of Theorem 2 for a Double Root**

A general solution of $y'' - 2y' + y = 0$ on any interval is $y = (c_1 + c_2 x)e^x$. (Verify!). The corresponding Wronskian is not 0, which shows linear independence of e^x and xe^x on any interval. Namely,

$$W(x, xe^x) = \begin{vmatrix} e^x & xe^x \\ e^x & (x+1)e^x \end{vmatrix} = (x+1)e^{2x} - xe^{2x} = e^{2x} \neq 0.$$ ■

A General Solution of (1) Includes All Solutions

This will be our second main result, as announced at the beginning. Let us start with existence.

THEOREM 3

> **Existence of a General Solution**
>
> *If $p(x)$ and $q(x)$ are continuous on an open interval I, then* (1) *has a general solution on I.*

PROOF By Theorem 1, the ODE (1) has a solution $y_1(x)$ on I satisfying the initial conditions

$$y_1(x_0) = 1, \qquad y_1'(x_0) = 0$$

and a solution $y_2(x)$ on I satisfying the initial conditions

$$y_2(x_0) = 0, \qquad y_2'(x_0) = 1.$$

The Wronskian of these two solutions has at $x = x_0$ the value

$$W(y_1(0), y_2(0)) = y_1(x_0)y_2'(x_0) - y_2(x_0)y_1'(x_0) = 1.$$

Hence, by Theorem 2, these solutions are linearly independent on I. They form a basis of solutions of (1) on I, and $y = c_1 y_1 + c_2 y_2$ with arbitrary c_1, c_2 is a general solution of (1) on I, whose existence we wanted to prove. ■

We finally show that a general solution is as general as it can possibly be.

THEOREM 4

A General Solution Includes All Solutions

If the ODE (1) *has continuous coefficients* $p(x)$ *and* $q(x)$ *on some open interval* I, *then every solution* $y = Y(x)$ *of* (1) *on* I *is of the form*

$$Y(x) = C_1 y_1(x) + C_2 y_2(x) \tag{8}$$

where y_1, y_2 *is any basis of solutions of* (1) *on* I *and* C_1, C_2 *are suitable constants.*

Hence (1) *does not have* **singular solutions** *(that is, solutions not obtainable from a general solution).*

PROOF Let $y = Y(x)$ be any solution of (1) on I. Now, by Theorem 3 the ODE (1) has a general solution

$$y(x) = c_1 y_1(x) + c_2 y_2(x) \tag{9}$$

on I. We have to find suitable values of c_1, c_2 such that $y(x) = Y(x)$ on I. We choose any x_0 in I and show first that we can find values of c_1, c_2 such that we reach agreement at x_0, that is, $y(x_0) = Y(x_0)$ and $y'(x_0) = Y'(x_0)$. Written out in terms of (9), this becomes

$$\text{(10)} \qquad \begin{aligned} &\text{(a)} \quad c_1 y_1(x_0) + c_2 y_2(x_0) = Y(x_0) \\ &\text{(b)} \quad c_1 y_1'(x_0) + c_2 y_2'(x_0) = Y'(x_0). \end{aligned}$$

We determine the unknowns c_1 and c_2. To eliminate c_2, we multiply (10a) by $y_2'(x_0)$ and (10b) by $-y_2(x_0)$ and add the resulting equations. This gives an equation for c_1. Then we multiply (10a) by $-y_1'(x_0)$ and (10b) by $y_1(x_0)$ and add the resulting equations. This gives an equation for c_2. These new equations are as follows, where we take the values of $y_1, y_1', y_2, y_2', Y, Y'$ at x_0.

$$c_1(y_1 y_2' - y_2 y_1') = c_1 W(y_1, y_2) = Y y_2' - y_2 Y'$$
$$c_2(y_1 y_2' - y_2 y_1') = c_2 W(y_1, y_2) = y_1 Y' - Y y_1'.$$

Since y_1, y_2 is a basis, the Wronskian W in these equations is not 0, and we can solve for c_1 and c_2. We call the (unique) solution $c_1 = C_1, c_2 = C_2$. By substituting it into (9) we obtain from (9) the particular solution

$$y^*(x) = C_1 y_1(x) + C_2 y_2(x).$$

Now since C_1, C_2 is a solution of (10), we see from (10) that

$$y^*(x_0) = Y(x_0), \qquad y^{*\prime}(x_0) = Y'(x_0).$$

From the uniqueness stated in Theorem 1 this implies that y^* and Y must be equal everywhere on I, and the proof is complete. ■

Reflecting on this section, we note that homogeneous linear ODEs with continuous variable coefficients have a conceptually and structurally rather transparent existence and uniqueness theory of solutions. Important in itself, this theory will also provide the foundation for our study of nonhomogeneous linear ODEs, whose theory and engineering applications form the content of the remaining four sections of this chapter.

PROBLEM SET 2.6

1. Derive (6*) from (6).

2–8 BASIS OF SOLUTIONS. WRONSKIAN

Find the Wronskian. Show linear independence by using quotients and confirm it by Theorem 2.

2. $e^{4.0x}, e^{-1.5x}$
3. $e^{-0.4x}, e^{-2.6x}$
4. $x, 1/x$
5. x^3, x^2
6. $e^{-x} \cos \omega x, e^{-x} \sin \omega x$
7. $\cosh ax, \sinh ax$
8. $x^k \cos (\ln x), x^k \sin (\ln x)$

9–15 ODE FOR GIVEN BASIS. WRONSKIAN. IVP

(a) Find a second-order homogeneous linear ODE for which the given functions are solutions. **(b)** Show linear independence by the Wronskian. **(c)** Solve the initial value problem.

9. $\cos 5x, \sin 5x, \quad y(0) = 3, \quad y'(0) = -5$
10. $x^{m_1}, x^{m_2}, \quad y(1) = -2, \quad y'(1) = 2m_1 - 4m_2$
11. $e^{-2.5x} \cos 0.3x, e^{-2.5x} \sin 0.3x, \quad y(0) = 3, \quad y'(0) = -7.5$
12. $x^2, x^2 \ln x, \quad y(1) = 4, \quad y'(1) = 6$
13. $1, e^{-2x}, \quad y(0) = 1, \quad y'(0) = -1$
14. $e^{-kx} \cos \pi x, e^{-kx} \sin \pi x, \quad y(0) = 1, \quad y'(0) = -k - \pi$
15. $\cosh 1.8x, \sinh 1.8x, \quad y(0) = 14.20, \quad y'(0) = 16.38$

16. **TEAM PROJECT. Consequences of the Present Theory.** This concerns some noteworthy general properties of solutions. Assume that the coefficients p and q of the ODE (1) are continuous on some open interval I, to which the subsequent statements refer.

(a) Solve $y'' - y = 0$ (a) by exponential functions, (b) by hyperbolic functions. How are the constants in the corresponding general solutions related?

(b) Prove that the solutions of a basis cannot be 0 at the same point.

(c) Prove that the solutions of a basis cannot have a maximum or minimum at the same point.

(d) Why is it likely that formulas of the form (6*) should exist?

(e) Sketch $y_1(x) = x^3$ if $x \geqq 0$ and 0 if $x < 0$, $y_2(x) = 0$ if $x \geqq 0$ and x^3 if $x < 0$. Show linear independence on $-1 < x < 1$. What is their Wronskian? What Euler–Cauchy equation do y_1, y_2 satisfy? Is there a contradiction to Theorem 2?

(f) Prove **Abel's formula**[6]

$$W(y_1(x), y_2(x)) = c \exp \left[-\int_{x_0}^{x} p(t)\, dt \right]$$

where $c = W(y_1(x_0), y_2(x_0))$. Apply it to Prob. 6. *Hint:* Write (1) for y_1 and for y_2. Eliminate q algebraically from these two ODEs, obtaining a first-order linear ODE. Solve it.

2.7 Nonhomogeneous ODEs

We now advance from homogeneous to nonhomogeneous linear ODEs.

Consider the second-order nonhomogeneous linear ODE

$$y'' + p(x)y' + q(x)y = r(x) \tag{1}$$

where $r(x) \not\equiv 0$. We shall see that a "general solution" of (1) is the sum of a general solution of the corresponding homogeneous ODE

[6]**NIELS HENRIK ABEL** (1802–1829), Norwegian mathematician.

$$y'' + p(x)y' + q(x)y = 0 \tag{2}$$

and a "particular solution" of (1). These two new terms "general solution of (1)" and "particular solution of (1)" are defined as follows.

DEFINITION

General Solution, Particular Solution

A **general solution** of the nonhomogeneous ODE (1) on an open interval I is a solution of the form

$$y(x) = y_h(x) + y_p(x); \tag{3}$$

here, $y_h = c_1y_1 + c_2y_2$ is a general solution of the homogeneous ODE (2) on I and y_p is any solution of (1) on I containing no arbitrary constants.

A **particular solution** of (1) on I is a solution obtained from (3) by assigning specific values to the arbitrary constants c_1 and c_2 in y_h.

Our task is now twofold, first to justify these definitions and then to develop a method for finding a solution y_p of (1).

Accordingly, we first show that a general solution as just defined satisfies (1) and that the solutions of (1) and (2) are related in a very simple way.

THEOREM 1

Relations of Solutions of (1) to Those of (2)

(a) *The sum of a solution y of* (1) *on some open interval I and a solution $\tilde{y}$ of* (2) *on I is a solution of* (1) *on I. In particular,* (3) *is a solution of* (1) *on I.*

(b) *The difference of two solutions of* (1) *on I is a solution of* (2) *on I.*

PROOF **(a)** Let $L[y]$ denote the left side of (1). Then for any solutions y of (1) and $\tilde{y}$ of (2) on I,

$$L[y + \tilde{y}] = L[y] + L[\tilde{y}] = r + 0 = r.$$

(b) For any solutions y and y^* of (1) on I we have $L[y - y^*] = L[y] - L[y^*] = r - r = 0$. ■

Now for *homogeneous ODEs* (2) we know that general solutions include all solutions. We show that the same is true for nonhomogeneous ODEs (1).

THEOREM 2

A General Solution of a Nonhomogeneous ODE Includes All Solutions

If the coefficients $p(x)$, $q(x)$, and the function $r(x)$ in (1) *are continuous on some open interval I, then every solution of* (1) *on I is obtained by assigning suitable values to the arbitrary constants c_1 and c_2 in a general solution* (3) *of* (1) *on I.*

PROOF Let y^* be any solution of (1) on I and x_0 any x in I. Let (3) be any general solution of (1) on I. This solution exists. Indeed, $y_h = c_1y_1 + c_2y_2$ exists by Theorem 3 in Sec. 2.6

because of the continuity assumption, and y_p exists according to a construction to be shown in Sec. 2.10. Now, by Theorem 1(b) just proved, the difference $Y = y^* - y_p$ is a solution of (2) on I. At x_0 we have

$$Y(x_0) = y^*(x_0) - y_p(x_0). \qquad Y'(x_0) = y^{*\prime}(x_0) - y_p'(x_0).$$

Theorem 1 in Sec. 2.6 implies that for these conditions, as for any other initial conditions in I, there exists a unique particular solution of (2) obtained by assigning suitable values to c_1, c_2 in y_h. From this and $y^* = Y + y_p$ the statement follows. ■

Method of Undetermined Coefficients

Our discussion suggests the following. *To solve the nonhomogeneous ODE* (1) *or an initial value problem for* (1), *we have to solve the homogeneous ODE* (2) *and find any solution* y_p *of* (1), so that we obtain a general solution (3) of (1).

How can we find a solution y_p of (1)? One method is the so-called **method of undetermined coefficients**. It is much simpler than another, more general, method (given in Sec. 2.10). Since it applies to models of vibrational systems and electric circuits to be shown in the next two sections, it is frequently used in engineering.

More precisely, the method of undetermined coefficients is suitable for linear ODEs with ***constant coefficients a and b***

$$y'' + ay' + by = r(x) \tag{4}$$

when $r(x)$ is an exponential function, a power of x, a cosine or sine, or sums or products of such functions. These functions have derivatives similar to $r(x)$ itself. This gives the idea. We choose a form for y_p similar to $r(x)$, but with unknown coefficients to be determined by substituting that y_p and its derivatives into the ODE. Table 2.1 on p. 82 shows the choice of y_p for practically important forms of $r(x)$. Corresponding rules are as follows.

Choice Rules for the Method of Undetermined Coefficients

(a) Basic Rule. *If $r(x)$ in* (4) *is one of the functions in the first column in Table* 2.1, *choose y_p in the same line and determine its undetermined coefficients by substituting y_p and its derivatives into* (4).

(b) Modification Rule. *If a term in your choice for y_p happens to be a solution of the homogeneous ODE corresponding to* (4), *multiply this term by x* (*or by x^2 if this solution corresponds to a double root of the characteristic equation of the homogeneous ODE*).

(c) Sum Rule. *If $r(x)$ is a sum of functions in the first column of Table* 2.1, *choose for y_p the sum of the functions in the corresponding lines of the second column.*

The Basic Rule applies when $r(x)$ is a single term. The Modification Rule helps in the indicated case, and to recognize such a case, we have to solve the homogeneous ODE first. The Sum Rule follows by noting that the sum of two solutions of (1) with $r = r_1$ and $r = r_2$ (and the same left side!) is a solution of (1) with $r = r_1 + r_2$. (Verify!)

The method is self-correcting. A false choice for y_p or one with too few terms will lead to a contradiction. A choice with too many terms will give a correct result, with superfluous coefficients coming out zero.

Let us illustrate Rules (a)–(c) by the typical Examples 1–3.

Table 2.1 Method of Undetermined Coefficients

Term in $r(x)$	Choice for $y_p(x)$
$ke^{\gamma x}$	$Ce^{\gamma x}$
$kx^n \ (n = 0, 1, \cdots)$	$K_n x^n + K_{n-1}x^{n-1} + \cdots + K_1 x + K_0$
$k \cos \omega x$	$K \cos \omega x + M \sin \omega x$
$k \sin \omega x$	
$ke^{\alpha x} \cos \omega x$	$e^{\alpha x}(K \cos \omega x + M \sin \omega x)$
$ke^{\alpha x} \sin \omega x$	

EXAMPLE 1 Application of the Basic Rule (a)

Solve the initial value problem

$$y'' + y = 0.001x^2, \qquad y(0) = 0, \qquad y'(0) = 1.5. \tag{5}$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The ODE $y'' + y = 0$ has the general solution

$$y_h = A \cos x + B \sin x.$$

Step 2. Solution y_p of the nonhomogeneous ODE. We first try $y_p = Kx^2$. Then $y_p'' = 2K$. By substitution, $2K + Kx^2 = 0.001x^2$. For this to hold for all x, the coefficient of each power of x (x^2 and x^0) must be the same on both sides; thus $K = 0.001$ and $2K = 0$, a contradiction.

The second line in Table 2.1 suggests the choice

$$y_p = K_2x^2 + K_1x + K_0. \qquad \text{Then} \qquad y_p'' + y_p = 2K_2 + K_2x^2 + K_1x + K_0 = 0.001x^2.$$

Equating the coefficients of x^2, x, x^0 on both sides, we have $K_2 = 0.001, K_1 = 0, 2K_2 + K_0 = 0$. Hence $K_0 = -2K_2 = -0.002$. This gives $y_p = 0.001x^2 - 0.002$, and

$$y = y_h + y_p = A \cos x + B \sin x + 0.001x^2 - 0.002.$$

Step 3. Solution of the initial value problem. Setting $x = 0$ and using the first initial condition gives $y(0) = A - 0.002 = 0$, hence $A = 0.002$. By differentiation and from the second initial condition,

$$y' = y_h' + y_p' = -A \sin x + B \cos x + 0.002x \qquad \text{and} \qquad y'(0) = B = 1.5.$$

This gives the answer (Fig. 50)

$$y = 0.002 \cos x + 1.5 \sin x + 0.001x^2 - 0.002.$$

Figure 50 shows y as well as the quadratic parabola y_p about which y is oscillating, practically like a sine curve since the cosine term is smaller by a factor of about $1/1000$. ■

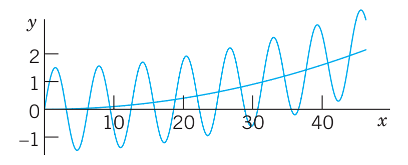


Fig. 50. Solution in Example 1

EXAMPLE 2 **Application of the Modification Rule (b)**

Solve the initial value problem

$$y'' + 3y' + 2.25y = -10e^{-1.5x}, \qquad y(0) = 1, \qquad y'(0) = 0. \tag{6}$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The characteristic equation of the homogeneous ODE is $\lambda^2 + 3\lambda + 2.25 = (\lambda + 1.5)^2 = 0$. Hence the homogeneous ODE has the general solution

$$y_h = (c_1 + c_2x)e^{-1.5x}.$$

Step 2. Solution y_p of the nonhomogeneous ODE. The function $e^{-1.5x}$ on the right would normally require the choice $Ce^{-1.5x}$. But we see from y_h that this function is a solution of the homogeneous ODE, which corresponds to a *double root* of the characteristic equation. Hence, according to the Modification Rule we have to multiply our choice function by x^2. That is, we choose

$$y_p = Cx^2e^{-1.5x}. \qquad \text{Then} \qquad y_p' = C(2x - 1.5x^2)e^{-1.5x}, \qquad y_p'' = C(2 - 3x - 3x + 2.25x^2)e^{-1.5x}.$$

We substitute these expressions into the given ODE and omit the factor $e^{-1.5x}$. This yields

$$C(2 - 6x + 2.25x^2) + 3C(2x - 1.5x^2) + 2.25Cx^2 = -10.$$

Comparing the coefficients of x^2, x, x^0 gives $0 = 0, 0 = 0, 2C = -10$, hence $C = -5$. This gives the solution $y_p = -5x^2e^{-1.5x}$. Hence the given ODE has the general solution

$$y = y_h + y_p = (c_1 + c_2x)e^{-1.5x} - 5x^2e^{-1.5x}.$$

Step 3. Solution of the initial value problem. Setting $x = 0$ in y and using the first initial condition, we obtain $y(0) = c_1 = 1$. Differentiation of y gives

$$y' = (c_2 - 1.5c_1 - 1.5c_2x)e^{-1.5x} - 10xe^{-1.5x} + 7.5x^2e^{-1.5x}.$$

From this and the second initial condition we have $y'(0) = c_2 - 1.5c_1 = 0$. Hence $c_2 = 1.5c_1 = 1.5$. This gives the answer (Fig. 51)

$$y = (1 + 1.5x)e^{-1.5x} - 5x^2e^{-1.5x} = (1 + 1.5x - 5x^2)e^{-1.5x}.$$

The curve begins with a horizontal tangent, crosses the x-axis at $x = 0.6217$ (where $1 + 1.5x - 5x^2 = 0$) and approaches the axis from below as x increases. ■

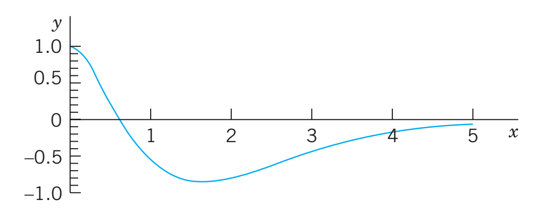


Fig. 51. Solution in Example 2

EXAMPLE 3 **Application of the Sum Rule (c)**

Solve the initial value problem

$$y'' + 2y' + 0.75y = 2\cos x - 0.25\sin x + 0.09x, \qquad y(0) = 2.78, \qquad y'(0) = -0.43. \tag{7}$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** The characteristic equation of the homogeneous ODE is

$$\lambda^2 + 2\lambda + 0.75 = (\lambda + \tfrac{1}{2})(\lambda + \tfrac{3}{2}) = 0$$

which gives the general solution $y_h = c_1e^{-x/2} + c_2e^{-3x/2}$.

Step 2. Particular solution of the nonhomogeneous ODE. We write $y_p = y_{p1} + y_{p2}$ and, following Table 2.1, (C) and (B),

$$y_{p1} = K \cos x + M \sin x \qquad \text{and} \qquad y_{p2} = K_1 x + K_0.$$

Differentiation gives $y'_{p1} = -K \sin x + M \cos x$, $y''_{p1} = -K \cos x - M \sin x$ and $y'_{p2} = 1$, $y''_{p2} = 0$. Substitution of y_{p1} into the ODE in (7) gives, by comparing the cosine and sine terms,

$$-K + 2M + 0.75K = 2, \qquad -M - 2K + 0.75M = -0.25,$$

hence $K = 0$ and $M = 1$. Substituting y_{p2} into the ODE in (7) and comparing the x- and x^0-terms gives

$$0.75K_1 = 0.09, \quad 2K_1 + 0.75K_0 = 0, \qquad \text{thus} \qquad K_1 = 0.12, \quad K_0 = -0.32.$$

Hence a general solution of the ODE in (7) is

$$y = c_1 e^{-x/2} + c_2 e^{-3x/2} + \sin x + 0.12x - 0.32.$$

Step 3. Solution of the initial value problem. From y, y' and the initial conditions we obtain

$$y(0) = c_1 + c_2 - 0.32 = 2.78, \qquad y'(0) = -\tfrac{1}{2}c_1 - \tfrac{3}{2}c_2 + 1 + 0.12 = -0.4.$$

Hence $c_1 = 3.1$, $c_2 = 0$. This gives the solution of the IVP (Fig. 52)

$$y = 3.1e^{-x/2} + \sin x + 0.12x - 0.32.$$

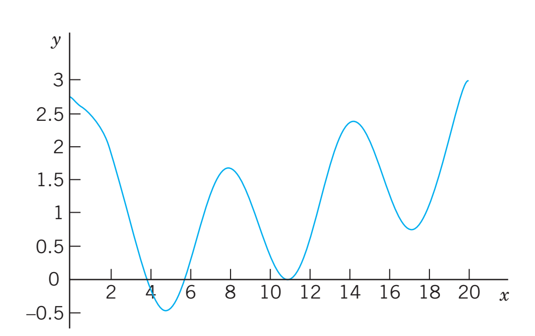


Fig. 52. Solution in Example 3

Stability. The following is important. If (and only if) all the roots of the characteristic equation of the homogeneous ODE $y'' + ay' + by = 0$ in (4) are negative, or have a negative real part, then a general solution y_h of this ODE goes to 0 as $x \to \infty$, so that the "**transient solution**" $y = y_h + y_p$ of (4) approaches the "**steady-state solution**" y_p. In this case the nonhomogeneous ODE and the physical or other system modeled by the ODE are called **stable**; otherwise they are called **unstable**. For instance, the ODE in Example 1 is unstable.

Applications follow in the next two sections.

PROBLEM SET 2.7

1–10 NONHOMOGENEOUS LINEAR ODEs: GENERAL SOLUTION

Find a (real) general solution. State which rule you are using. Show each step of your work.

1. $y'' + 5y' + 4y = 10e^{-3x}$

2. $10y'' + 50y' + 57.6y = \cos x$

3. $y'' + 3y' + 2y = 12x^2$

4. $y'' - 9y = 18 \cos \pi x$

5. $y'' + 4y' + 4y = e^{-x} \cos x$

6. $y'' + y' + (\pi^2 + \tfrac{1}{4})y = e^{-x/2} \sin \pi x$

7. $(D^2 + 2D + \frac{3}{4}I)y = 3e^x + \frac{9}{2}x$

8. $(3D^2 + 27I)y = 3 \cos x + \cos 3x$

9. $(D^2 - 16I)y = 9.6e^{4x} + 30e^x$

10. $(D^2 + 2D + I)y = 2x \sin x$

11–18 NONHOMOGENEOUS LINEAR ODEs: IVPs

Solve the initial value problem. State which rule you are using. Show each step of your calculation in detail.

11. $y'' + 3y = 18x^2, \quad y(0) = -3, \quad y'(0) = 0$

12. $y'' + 4y = -12 \sin 2x, \quad y(0) = 1.8, \quad y'(0) = 5.0$

13. $8y'' - 6y' + y = 6 \cosh x, \quad y(0) = 0.2,$ $y'(0) = 0.05$

14. $y'' + 4y' + 4y = e^{-2x} \sin 2x, \quad y(0) = 1,$ $y'(0) = -1.5$

15. $(x^2D^2 - 3xD + 3I)y = 3 \ln x - 4,$ $y(1) = 0, \quad y'(1) = 1; \quad y_p = \ln x$

16. $(D^2 - 2D)y = 6e^{2x} - 4e^{-2x}, \quad y(0) = -1, \quad y'(0) = 6$

17. $(D^2 + 0.2D + 0.26I)y = 1.22e^{0.5x}, \quad y(0) = 3.5,$ $y'(0) = 0.35$

18. $(D^2 + 2D + 10I)y = 17 \sin x - 37 \sin 3x,$ $y(0) = 6.6, \quad y'(0) = -2.2$

19. CAS PROJECT. Structure of Solutions of Initial Value Problems. Using the present method, find, graph, and discuss the solutions y of initial value problems of your own choice. Explore effects on solutions caused by changes of initial conditions. Graph $y_p, y, y - y_p$ separately, to see the separate effects. Find a problem in which **(a)** the part of y resulting from y_h decreases to zero, **(b)** increases, **(c)** is not present in the answer y. Study a problem with $y(0) = 0, y'(0) = 0$. Consider a problem in which you need the Modification Rule (a) for a simple root, (b) for a double root. Make sure that your problems cover all three Cases I, II, III (see Sec. 2.2).

20. TEAM PROJECT. Extensions of the Method of Undetermined Coefficients. (a) Extend the method to products of the function in Table 2.1, **(b)** Extend the method to Euler–Cauchy equations. Comment on the practical significance of such extensions.

2.8 Modeling: Forced Oscillations. Resonance

In Sec. 2.4 we considered vertical motions of a mass–spring system (vibration of a mass m on an elastic spring, as in Figs. 33 and 53) and modeled it by the ***homogeneous*** linear ODE

$$my'' + cy' + ky = 0. \tag{1}$$

Here $y(t)$ as a function of time t is the displacement of the body of mass m from rest.

The mass–spring system of Sec. 2.4 exhibited only free motion. This means no external forces (outside forces) but only internal forces controlled the motion. The internal forces are forces within the system. They are the force of inertia my'', the damping force cy' (if $c > 0$), and the spring force ky, a restoring force.

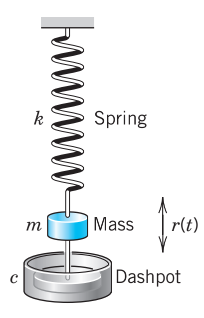


Fig. 53. Mass on a spring

We now extend our model by including an additional force, that is, the external force $r(t)$, on the right. Then we have

$$(2^*) \qquad my'' + cy' + ky = r(t).$$

Mechanically this means that at each instant t the resultant of the internal forces is in equilibrium with $r(t)$. The resulting motion is called a **forced motion** with **forcing function** $r(t)$, which is also known as **input** or **driving force**, and the solution $y(t)$ to be obtained is called the **output** or the **response** *of the system to the driving force.*

Of special interest are periodic external forces, and we shall consider a driving force of the form

$$r(t) = F_0 \cos \omega t \qquad (F_0 > 0, \omega > 0).$$

Then we have the nonhomogeneous ODE

$$\textbf{(2)} \qquad my'' + cy' + ky = F_0 \cos \omega t.$$

Its solution will reveal facts that are fundamental in engineering mathematics and allow us to model resonance.

Solving the Nonhomogeneous ODE (2)

From Sec. 2.7 we know that a general solution of (2) is the sum of a general solution y_h of the homogeneous ODE (1) plus any solution y_p of (2). To find y_p, we use the method of undetermined coefficients (Sec. 2.7), starting from

$$\textbf{(3)} \qquad y_p(t) = a \cos \omega t + b \sin \omega t.$$

By differentiating this function (chain rule!) we obtain

$$y_p' = -\omega a \sin \omega t + \omega b \cos \omega t,$$
$$y_p'' = -\omega^2 a \cos \omega t - \omega^2 b \sin \omega t.$$

Substituting y_p, y_p', and y_p'' into (2) and collecting the cosine and the sine terms, we get

$$[(k - m\omega^2)a + \omega cb] \cos \omega t + [-\omega ca + (k - m\omega^2)b] \sin \omega t = F_0 \cos \omega t.$$

The cosine terms on both sides must be equal, and the coefficient of the sine term on the left must be zero since there is no sine term on the right. This gives the two equations

$$(4) \qquad \begin{aligned} (k - m\omega^2)a + \quad \omega cb \quad &= F_0 \\ -\omega ca \quad + (k - m\omega^2)b &= 0 \end{aligned}$$

for determining the unknown coefficients a and b. This is a linear system. We can solve it by elimination. To eliminate b, multiply the first equation by $k - m\omega^2$ and the second by $-\omega c$ and add the results, obtaining

$$(k - m\omega^2)^2 a + \omega^2 c^2 a = F_0(k - m\omega^2).$$

Similarly, to eliminate a, multiply (the first equation by ωc and the second by $k - m\omega^2$ and add to get

$$\omega^2 c^2 b + (k - m\omega^2)^2 b = F_0 \omega c.$$

If the factor $(k - m\omega^2)^2 + \omega^2 c^2$ is not zero, we can divide by this factor and solve for a and b,

$$a = F_0 \frac{k - m\omega^2}{(k - m\omega^2)^2 + \omega^2 c^2}, \qquad b = F_0 \frac{\omega c}{(k - m\omega^2)^2 + \omega^2 c^2}.$$

If we set $\sqrt{k/m} = \omega_0 \; (> 0)$ as in Sec. 2.4, then $k = m\omega_0^2$ and we obtain

(5) $$a = F_0 \frac{m(\omega_0^2 - \omega^2)}{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}, \qquad b = F_0 \frac{\omega c}{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}.$$

We thus obtain the general solution of the nonhomogeneous ODE (2) in the form

(6) $$y(t) = y_h(t) + y_p(t).$$

Here y_h is a general solution of the homogeneous ODE (1) and y_p is given by (3) with coefficients (5).

We shall now discuss the behavior of the mechanical system, distinguishing between the two cases $c = 0$ (no damping) and $c > 0$ (damping). These cases will correspond to two basically different types of output.

Case 1. Undamped Forced Oscillations. Resonance

If the damping of the physical system is so small that its effect can be neglected over the time interval considered, we can set $c = 0$. Then (5) reduces to $a = F_0/[m(\omega_0^2 - \omega^2)]$ and $b = 0$. Hence (3) becomes (use $\omega_0{}^2 = k/m$)

(7) $$y_p(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t = \frac{F_0}{k[1 - (\omega/\omega_0)^2]} \cos \omega t.$$

Here we must assume that $\omega^2 \neq \omega_0{}^2$; physically, the frequency $\omega/(2\pi)$ [cycles/sec] of the driving force is different from the *natural frequency* $\omega_0/(2\pi)$ of the system, which is the frequency of the free undamped motion [see (4) in Sec. 2.4]. From (7) and from (4*) in Sec. 2.4 we have the general solution of the "undamped system"

(8) $$y(t) = C \cos (\omega_0 t - \delta) + \frac{F_0}{m(\omega_0^2 - \omega^2)} \cos \omega t.$$

We see that this output is a ***superposition of two harmonic oscillations*** *of the frequencies just mentioned.*

Resonance. We discuss (7). We see that the maximum amplitude of y_p is (put $\cos \omega t = 1$)

$$a_0 = \frac{F_0}{k}\rho \qquad \text{where} \qquad \rho = \frac{1}{1 - (\omega/\omega_0)^2}. \tag{9}$$

a_0 depends on ω and ω_0. If $\omega \to \omega_0$, then ρ and a_0 tend to infinity. This excitation of large oscillations by matching input and natural frequencies ($\omega = \omega_0$) is called **resonance**. ρ is called the **resonance factor** (Fig. 54), and from (9) we see that $\rho/k = a_0/F_0$ is the ratio of the amplitudes of the particular solution y_p and of the input $F_0 \cos \omega t$. We shall see later in this section that resonance is of basic importance in the study of vibrating systems.

In the case of resonance the nonhomogeneous ODE (2) becomes

$$y'' + \omega_0^2 y = \frac{F_0}{m} \cos \omega_0 t. \tag{10}$$

Then (7) is no longer valid, and, from the Modification Rule in Sec. 2.7, we conclude that a particular solution of (10) is of the form

$$y_p(t) = t(a \cos \omega_0 t + b \sin \omega_0 t).$$

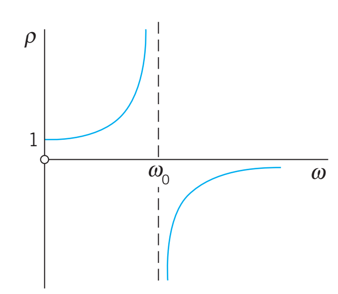


Fig. 54. Resonance factor $\rho(\omega)$

By substituting this into (10) we find $a = 0$ and $b = F_0/(2m\omega_0)$. Hence (Fig. 55)

$$y_p(t) = \frac{F_0}{2m\omega_0} t \sin \omega_0 t. \tag{11}$$

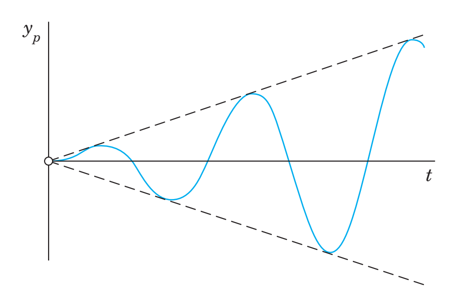


Fig. 55. Particular solution in the case of resonance

We see that, because of the factor t, the amplitude of the vibration becomes larger and larger. Practically speaking, systems with very little damping may undergo large vibrations

that can destroy the system. We shall return to this practical aspect of resonance later in this section.

Beats. Another interesting and highly important type of oscillation is obtained if ω is close to ω_0. Take, for example, the particular solution [see (8)]

$$y(t) = \frac{F_0}{m(\omega_0^2 - \omega^2)}(\cos \omega t - \cos \omega_0 t) \qquad (\omega \neq \omega_0). \tag{12}$$

Using (12) in App. 3.1, we may write this as

$$y(t) = \frac{2F_0}{m(\omega_0^2 - \omega^2)} \sin\left(\frac{\omega_0 + \omega}{2} t\right) \sin\left(\frac{\omega_0 - \omega}{2} t\right).$$

Since ω is close to ω_0, the difference $\omega_0 - \omega$ is small. Hence the period of the last sine function is large, and we obtain an oscillation of the type shown in Fig. 56, the dashed curve resulting from the first sine factor. This is what musicians are listening to when they ***tune*** their instruments.

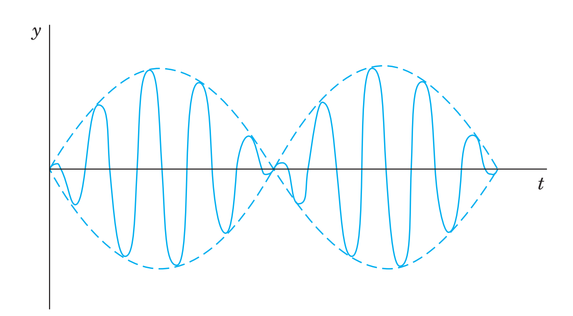


Fig. 56. Forced undamped oscillation when the difference of the input and natural frequencies is small (**"beats"**)

Case 2. Damped Forced Oscillations

If the damping of the mass–spring system is not negligibly small, we have $c > 0$ and a damping term cy' in (1) and (2). Then the general solution y_h of the homogeneous ODE (1) approaches zero as t goes to infinity, as we know from Sec. 2.4. Practically, it is zero after a sufficiently long time. Hence the "**transient solution**" (6) of (2), given by $y = y_h + y_p$, approaches the "**steady-state solution**" y_p. This proves the following.

THEOREM 1

Steady-State Solution

After a sufficiently long time the output of a damped vibrating system under a purely sinusoidal driving force [*see* (2)] *will practically be a harmonic oscillation whose frequency is that of the input.*

Amplitude of the Steady-State Solution. Practical Resonance

Whereas in the undamped case the amplitude of y_p approaches infinity as ω approaches ω_0, this will not happen in the damped case. In this case the amplitude will always be finite. But it may have a maximum for some ω depending on the damping constant c. This may be called **practical resonance**. It is of great importance because if c is not too large, then some input may excite oscillations large enough to damage or even destroy the system. Such cases happened, in particular in earlier times when less was known about resonance. Machines, cars, ships, airplanes, bridges, and high-rising buildings are vibrating mechanical systems, and it is sometimes rather difficult to find constructions that are completely free of undesired resonance effects, caused, for instance, by an engine or by strong winds.

To study the amplitude of y_p as a function of ω, we write (3) in the form

$$y_p(t) = C^* \cos(\omega t - \eta). \tag{13}$$

C^* is called the **amplitude** of y_p and η the **phase angle** or **phase lag** because it measures the lag of the output behind the input. According to (5), these quantities are

$$\begin{aligned} C^*(\omega) &= \sqrt{a^2 + b^2} = \frac{F_0}{\sqrt{m^2(\omega_0^2 - \omega^2)^2 + \omega^2 c^2}}, \\ \tan \eta(\omega) &= \frac{b}{a} = \frac{\omega c}{m(\omega_0^2 - \omega^2)}. \end{aligned} \tag{14}$$

Let us see whether $C^*(\omega)$ has a maximum and, if so, find its location and then its size. We denote the radicand in the second root in C^* by R. Equating the derivative of C^* to zero, we obtain

$$\frac{dC^*}{d\omega} = F_0\left(-\frac{1}{2}R^{-3/2}\right)[2m^2(\omega_0^2 - \omega^2)(-2\omega) + 2\omega c^2].$$

The expression in the brackets [. . .] is zero if

$$c^2 = 2m^2(\omega_0^2 - \omega^2) \qquad (\omega_0^2 = k/m). \tag{15}$$

By reshuffling terms we have

$$2m^2\omega^2 = 2m^2{\omega_0}^2 - c^2 = 2mk - c^2.$$

The right side of this equation becomes negative if $c^2 > 2mk$, so that then (15) has no real solution and C^* decreases monotone as ω increases, as the lowest curve in Fig. 57 shows. If c is smaller, $c^2 < 2mk$, then (15) has a real solution $\omega = \omega_{\max}$, where

$$\omega_{\max}^2 = \omega_0^2 - \frac{c^2}{2m^2}. \tag{15*}$$

From (15*) we see that this solution increases as c decreases and approaches ω_0 as c approaches zero. See also Fig. 57.

The size of $C^*(\omega_{\max})$ is obtained from (14), with $\omega^2 = \omega^2_{\max}$ given by (15*). For this ω^2 we obtain in the second radicand in (14) from (15*)

$$m^2(\omega_0^2 - \omega^2_{\max})^2 = \frac{c^4}{4m^2} \qquad \text{and} \qquad \omega^2_{\max} c^2 = \left(\omega_0^2 - \frac{c^2}{2m^2}\right)c^2.$$

The sum of the right sides of these two formulas is

$$(c^4 + 4m^2\omega_0^2 c^2 - 2c^4)/(4m^2) = c^2(4m^2\omega_0^2 - c^2)/(4m^2).$$

Substitution into (14) gives

$$C^*(\omega_{\max}) = \frac{2mF_0}{c\sqrt{4m^2\omega_0^2 - c^2}}. \tag{16}$$

We see that $C^*(\omega_{\max})$ is always finite when $c > 0$. Furthermore, since the expression

$$c^2 4m^2\omega_0^2 - c^4 = c^2(4mk - c^2)$$

in the denominator of (16) decreases monotone to zero as c^2 $(<2mk)$ goes to zero, the maximum amplitude (16) increases monotone to infinity, in agreement with our result in Case 1. Figure 57 shows the **amplification** C^*/F_0 (ratio of the amplitudes of output and input) as a function of ω for $m = 1$, $k = 1$, hence $\omega_0 = 1$, and various values of the damping constant c.

Figure 58 shows the phase angle (the lag of the output behind the input), which is less than $\pi/2$ when $\omega < \omega_0$, and greater than $\pi/2$ for $\omega > \omega_0$.

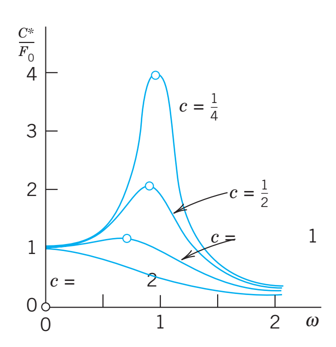


Fig. 57. Amplification C^*/F_0 as a function of ω for $m = 1$, $k = 1$, and various values of the damping constant c

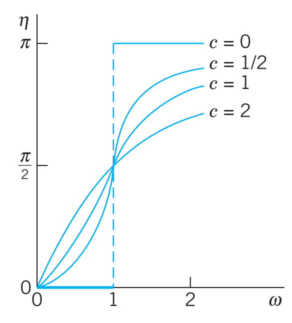


Fig. 58. Phase lag η as a function of ω for $m = 1$, $k = 1$, thus $\omega_0 = 1$, and various values of the damping constant c

PROBLEM SET 2.8

1. WRITING REPORT. Free and Forced Vibrations. Write a condensed report of 2–3 pages on the most important similarities and differences of free and forced vibrations, with examples of your own. No proofs.

2. Which of Probs. 1–18 in Sec. 2.7 (with x = time t) can be models of mass–spring systems with a harmonic oscillation as steady-state solution?

3–7 STEADY-STATE SOLUTIONS

Find the steady-state motion of the mass–spring system modeled by the ODE. Show the details of your work.

3. $y'' + 6y' + 8y = 42.5 \cos 2t$

4. $y'' + 2.5y' + 10y = -13.6 \sin 4t$

5. $(D^2 + D + 4.25I)y = 22.1 \cos 4.5t$

6. $(D^2 + 4D + 3I)y = \cos t + \frac{1}{3}\cos 3t$

7. $(4D^2 + 12D + 9I)y = 225 - 75 \sin 3t$

8–15 TRANSIENT SOLUTIONS

Find the transient motion of the mass–spring system modeled by the ODE. Show the details of your work.

8. $2y'' + 4y' + 6.5y = 4 \sin 1.5t$

9. $y'' + 3y' + 3.25y = 3 \cos t - 1.5 \sin t$

10. $y'' + 16y = 56 \cos 4t$

11. $(D^2 + 2I)y = \cos \sqrt{2}t + \sin \sqrt{2}t$

12. $(D^2 + 2D + 5I)y = 4 \cos t + 8 \sin t$

13. $(D^2 + I)y = \cos \omega t, \omega^2 \neq 1$

14. $(D^2 + I)y = 5e^{-t} \cos t$

15. $(D^2 + 4D + 8I)y = 2 \cos 2t + \sin 2t$

16–20 INITIAL VALUE PROBLEMS

Find the motion of the mass–spring system modeled by the ODE and the initial conditions. Sketch or graph the solution curve. In addition, sketch or graph the curve of $y - y_p$ to see when the system practically reaches the steady state.

16. $y'' + 25y = 24 \sin t, \quad y(0) = 1, \quad y'(0) = 1$

17. $(D^2 + 4I)y = \sin t + \frac{1}{3}\sin 3t + \frac{1}{5}\sin 5t$, $y(0) = 0, \quad y'(0) = \frac{3}{35}$

18. $(D^2 + 8D + 17I)y = 474.5 \sin 0.5t, \quad y(0) = -5.4$, $y'(0) = 9.4$

19. $(D^2 + 2D + 2I)y = e^{-t/2} \sin \frac{1}{2}t, \quad y(0) = 0$, $y'(0) = 1$

20. $(D^2 + 5I)y = \cos \pi t - \sin \pi t, \quad y(0) = 0, \quad y'(0) = 0$

21. Beats. Derive the formula after (12) from (12). Can we have beats in a damped system?

22. Beats. Solve $y'' + 25y = 99 \cos 4.9t$, $y(0) = 2$, $y'(0) = 0$. How does the graph of the solution change if you change **(a)** $y(0)$, **(b)** the frequency of the driving force?

23. TEAM EXPERIMENT. Practical Resonance.
(a) Derive, in detail, the crucial formula (16).

(b) By considering dC^*/dc show that $C^*(\omega_{max})$ increases as c ($\leqq \sqrt{2mk}$) decreases.

(c) Illustrate practical resonance with an ODE of your own in which you vary c, and sketch or graph corresponding curves as in Fig. 57.

(d) Take your ODE with c fixed and an input of two terms, one with frequency close to the practical resonance frequency and the other not. Discuss and sketch or graph the output.

(e) Give other applications (not in the book) in which resonance is important.

24. Gun barrel. Solve $y'' + y = 1 - t^2/\pi^2$ if $0 \leqq t \leqq \pi$ and 0 if $t \to \infty$; here, $y(0) = 0, y'(0) = 0$. This models an undamped system on which a force F acts during some interval of time (see Fig. 59), for instance, the force on a gun barrel when a shell is fired, the barrel being braked by heavy springs (and then damped by a dashpot, which we disregard for simplicity). *Hint:* At π both y and y' must be continuous.

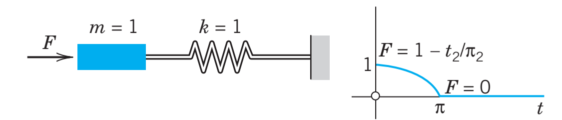
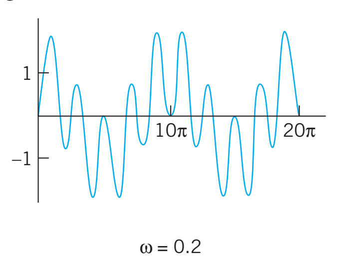


Fig. 59. Problem 24

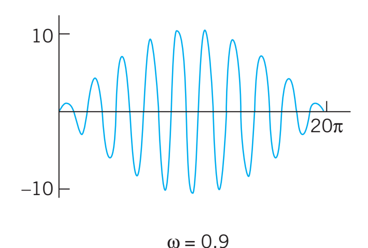
25. CAS EXPERIMENT. Undamped Vibrations.
(a) Solve the initial value problem $y'' + y = \cos \omega t$, $\omega^2 \neq 1$, $y(0) = 0, y'(0) = 0$. Show that the solution can be written

$$y(t) = \frac{2}{1 - \omega^2} \sin \left[\tfrac{1}{2}(1 + \omega)t\right] \sin \left[\tfrac{1}{2}(1 - \omega)t\right].$$

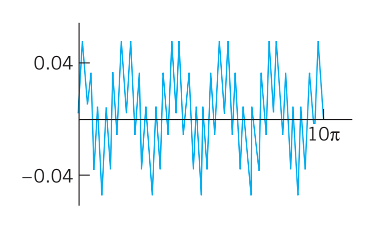
(b) Experiment with the solution by changing ω to see the change of the curves from those for small ω (>0) to beats, to resonance, and to large values of ω (see Fig. 60).



ω = 0.2



ω = 0.9



ω = 6

Fig. 60. Typical solution curves in CAS Experiment 25

2.9 Modeling: Electric Circuits

Designing good models is a task the computer cannot do. Hence setting up models has become an important task in modern applied mathematics. The best way to gain experience in successful modeling is to carefully examine the modeling process in various fields and applications. Accordingly, modeling electric circuits will be *profitable for all students,* not just for electrical engineers and computer scientists.

Figure 61 shows an ***RLC*-circuit**, as it occurs as a basic building block of large electric networks in computers and elsewhere. An *RLC*-circuit is obtained from an *RL*-circuit by adding a capacitor. Recall Example 2 on the *RL*-circuit in Sec. 1.5: The model of the *RL*-circuit is $LI' + RI = E(t)$. It was obtained by **KVL** (Kirchhoff's Voltage Law)[7] by equating the voltage drops across the resistor and the inductor to the EMF (electromotive force). Hence we obtain the model of the *RLC*-circuit simply by adding the voltage drop Q/C across the capacitor. Here, C F (farads) is the capacitance of the capacitor. Q coulombs is the charge on the capacitor, related to the current by

$$I(t) = \frac{dQ}{dt}, \qquad \text{equivalently} \qquad Q(t) = \int I(t)\, dt.$$

See also Fig. 62. Assuming a sinusoidal EMF as in Fig. 61, we thus have the model of the *RLC*-circuit

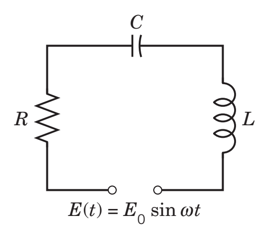


Fig. 61. *RLC*-circuit

Name	Symbol	Notation	Unit	Voltage Drop
Ohm's Resistor	(resistor)	R Ohm's Resistance	ohms (Ω)	RI
Inductor	(inductor)	L Inductance	henrys (H)	$L\dfrac{dI}{dt}$
Capacitor	(capacitor)	C Capacitance	farads (F)	Q/C

Fig. 62. Elements in an *RLC*-circuit

---

[7]GUSTAV ROBERT KIRCHHOFF (1824–1887), German physicist. Later we shall also need **Kirchhoff's Current Law (KCL)**:

At any point of a circuit, the sum of the inflowing currents is equal to the sum of the outflowing currents.

The units of measurement of electrical quantities are named after ANDRÉ MARIE AMPÈRE (1775–1836), French physicist, CHARLES AUGUSTIN DE COULOMB (1736–1806), French physicist and engineer, MICHAEL FARADAY (1791–1867), English physicist, JOSEPH HENRY (1797–1878), American physicist, GEORG SIMON OHM (1789–1854), German physicist, and ALESSANDRO VOLTA (1745–1827), Italian physicist.

$$(1') \qquad LI' + RI + \frac{1}{C}\int I\,dt = E(t) = E_0 \sin \omega t.$$

This is an "integro-differential equation." To get rid of the integral, we differentiate (1′) with respect to t, obtaining

$$\textbf{(1)} \qquad LI'' + RI' + \frac{1}{C} I = E'(t) = E_0 \omega \cos \omega t.$$

This shows that the current in an RLC-circuit is obtained as the solution of this nonhomogeneous second-order ODE (1) with constant coefficients.

In connection with initial value problems, we shall occasionally use

$$(1'') \qquad LQ'' + RQ'' + \frac{1}{C} Q = E(t),$$

obtained from (1′) and $I = Q'$.

Solving the ODE (1) for the Current in an *RLC*-Circuit

A general solution of (1) is the sum $I = I_h + I_p$, where I_h is a general solution of the homogeneous ODE corresponding to (1) and I_p is a particular solution of (1). We first determine I_p by the method of undetermined coefficients, proceeding as in the previous section. We substitute

$$(2) \qquad \begin{aligned} I_p &= a \cos \omega t + b \sin \omega t \\ I_p' &= \omega(-a \sin \omega t + b \cos \omega t) \\ I_p'' &= \omega^2(-a \cos \omega t - b \sin \omega t) \end{aligned}$$

into (1). Then we collect the cosine terms and equate them to $E_0 \omega \cos \omega t$ on the right, and we equate the sine terms to zero because there is no sine term on the right,

$$L\omega^2(-a) + R\omega b + a/C = E_0\omega \qquad \text{(Cosine terms)}$$

$$L\omega^2(-b) + R\omega(-a) + b/C = 0 \qquad \text{(Sine terms).}$$

Before solving this system for a and b, we first introduce a combination of L and C, called the **reactance**

$$\textbf{(3)} \qquad S = \omega L - \frac{1}{\omega C}.$$

Dividing the previous two equations by ω, ordering them, and substituting S gives

$$-Sa + Rb = E_0$$

$$-Ra - Sb = 0.$$

We now eliminate b by multiplying the first equation by S and the second by R, and adding. Then we eliminate a by multiplying the first equation by R and the second by $-S$, and adding. This gives

$$-(S^2 + R^2)a = E_0 S, \qquad (R^2 + S^2)b = E_0 R.$$

We can solve for a and b,

$$a = \frac{-E_0 S}{R^2 + S^2}, \qquad b = \frac{E_0 R}{R^2 + S^2}. \tag{4}$$

Equation (2) with coefficients a and b given by (4) is the desired particular solution I_p of the nonhomogeneous ODE (1) governing the current I in an RLC-circuit with sinusoidal electromotive force.

Using (4), we can write I_p in terms of "physically visible" quantities, namely, amplitude I_0 and phase lag θ of the current behind the EMF, that is,

$$I_p(t) = I_0 \sin(\omega t - \theta) \tag{5}$$

where [see (14) in App. A3.1]

$$I_0 = \sqrt{a^2 + b^2} = \frac{E_0}{\sqrt{R^2 + S^2}}, \qquad \tan\theta = -\frac{a}{b} = \frac{S}{R}.$$

The quantity $\sqrt{R^2 + S^2}$ is called the **impedance**. Our formula shows that the impedance equals the ratio E_0/I_0. This is somewhat analogous to $E/I = R$ (Ohm's law) and, because of this analogy, the impedance is also known as the **apparent resistance**.

A general solution of the homogeneous equation corresponding to (1) is

$$I_h = c_1 e^{\lambda_1 t} + c_2 e^{\lambda_2 t}$$

where λ_1 and λ_2 are the roots of the characteristic equation

$$\lambda^2 + \frac{R}{L}\lambda + \frac{1}{LC} = 0.$$

We can write these roots in the form $\lambda_1 = -\alpha + \beta$ and $\lambda_2 = -\alpha - \beta$, where

$$\alpha = \frac{R}{2L}, \qquad \beta = \sqrt{\frac{R^2}{4L^2} - \frac{1}{LC}} = \frac{1}{2L}\sqrt{R^2 - \frac{4L}{C}}.$$

Now in an actual circuit, R is never zero (hence $R > 0$). From this it follows that I_h approaches zero, theoretically as $t \to \infty$, but practically after a relatively short time. Hence the transient current $I = I_h + I_p$ tends to the steady-state current I_p, and after some time the output will practically be a harmonic oscillation, which is given by (5) and whose frequency is that of the input (of the electromotive force).

EXAMPLE 1 ***RLC*-Circuit**

Find the current $I(t)$ in an RLC-circuit with $R = 11\ \Omega$ (ohms), $L = 0.1$ H (henry), $C = 10^{-2}$ F (farad), which is connected to a source of EMF $E(t) = 110 \sin (60 \cdot 2\pi t) = 110 \sin 377\, t$ (hence 60 Hz = 60 cycles/sec, the usual in the U.S. and Canada; in Europe it would be 220 V and 50 Hz). Assume that current and capacitor charge are 0 when $t = 0$.

Solution. ***Step 1. General solution of the homogeneous ODE.*** Substituting R, L, C and the derivative $E'(t)$ into (1), we obtain

$$0.1I'' + 11I' + 100I = 110 \cdot 377 \cos 377t.$$

Hence the homogeneous ODE is $0.1I'' + 11I' + 100I = 0$. Its characteristic equation is

$$0.1\lambda^2 + 11\lambda + 100 = 0.$$

The roots are $\lambda_1 = -10$ and $\lambda_2 = -100$. The corresponding general solution of the homogeneous ODE is

$$I_h(t) = c_1 e^{-10t} + c_2 e^{-100t}.$$

Step 2. Particular solution I_p of (1). We calculate the reactance $S = 37.7 - 0.3 = 37.4$ and the steady-state current

$$I_p(t) = a \cos 377t + b \sin 377t$$

with coefficients obtained from (4) (and rounded)

$$a = \frac{-110 \cdot 37.4}{11^2 + 37.4^2} = -2.71, \qquad b = \frac{110 \cdot 11}{11^2 + 37.4^2} = 0.796.$$

Hence in our present case, a general solution of the nonhomogeneous ODE (1) is

$$I(t) = c_1 e^{-10t} + c_2 e^{-100t} - 2.71 \cos 377t + 0.796 \sin 377t. \tag{6}$$

Step 3. Particular solution satisfying the initial conditions. **How to use $Q(0) = 0$?** We finally determine c_1 and c_2 from the in initial conditions $I(0) = 0$ and $Q(0) = 0$. From the first condition and (6) we have

$$I(0) = c_1 + c_2 - 2.71 = 0, \qquad \text{hence} \qquad c_2 = 2.71 - c_1. \tag{7}$$

We turn to $Q(0) = 0$. The integral in (1′) equals $\int I\, dt = Q(t)$; see near the beginning of this section. Hence for $t = 0$, Eq. (1′) becomes

$$LI'(0) + R \cdot 0 = 0, \qquad \text{so that} \qquad I'(0) = 0.$$

Differentiating (6) and setting $t = 0$, we thus obtain

$$I'(0) = -10c_1 - 100c_2 + 0 + 0.796 \cdot 377 = 0, \qquad \text{hence by (7),} \qquad -10c_1 = 100(2.71 - c_1) - 300.1.$$

The solution of this and (7) is $c_1 = -0.323$, $c_2 = 3.033$. Hence the answer is

$$I(t) = -0.323e^{-10t} + 3.033e^{-100t} - 2.71 \cos 377t + 0.796 \sin 377t.$$

You may get slightly different values depending on the rounding. Figure 63 shows $I(t)$ as well as $I_p(t)$, which practically coincide, except for a very short time near $t = 0$ because the exponential terms go to zero very rapidly. Thus after a very short time the current will practically execute harmonic oscillations of the input frequency 60 Hz = 60 cycles/sec. Its maximum amplitude and phase lag can be seen from (5), which here takes the form

$$I_p(t) = 2.824 \sin (377t - 1.29).$$ ■

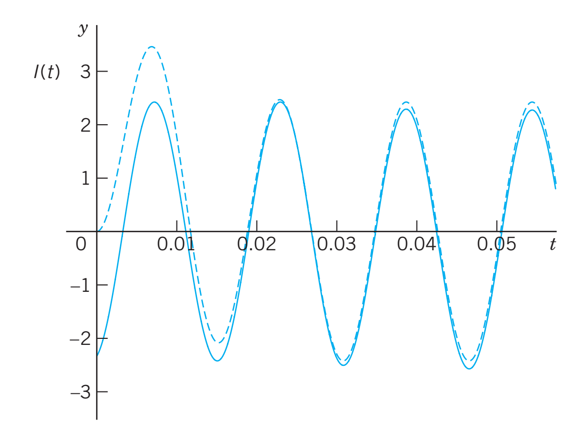


Fig. 63. Transient (upper curve) and steady-state currents in Example 1

Analogy of Electrical and Mechanical Quantities

Entirely different physical or other systems may have the same mathematical model. For instance, we have seen this from the various applications of the ODE $y' = ky$ in Chap. 1. Another impressive demonstration of this ***unifying power of mathematics*** is given by the ODE (1) for an electric RLC-circuit and the ODE (2) in the last section for a mass–spring system. Both equations

$$LI'' + RI' + \frac{1}{C}I = E_0\omega \cos \omega t \qquad \text{and} \qquad my'' + cy' + ky = F_0 \cos \omega t$$

are of the same form. Table 2.2 shows the analogy between the various quantities involved. The inductance L corresponds to the mass m and, indeed, an inductor opposes a change in current, having an "inertia effect" similar to that of a mass. The resistance R corresponds to the damping constant c, and a resistor causes loss of energy, just as a damping dashpot does. And so on.

This analogy is ***strictly quantitative*** in the sense that to a given mechanical system we can construct an electric circuit whose current will give the exact values of the displacement in the mechanical system when suitable scale factors are introduced.

The ***practical importance*** of this analogy is almost obvious. The analogy may be used for constructing an "electrical model" of a given mechanical model, resulting in substantial savings of time and money because electric circuits are easy to assemble, and electric quantities can be measured much more quickly and accurately than mechanical ones.

Table 2.2 Analogy of Electrical and Mechanical Quantities

Electrical System	Mechanical System
Inductance L	Mass m
Resistance R	Damping constant c
Reciprocal $1/C$ of capacitance	Spring modulus k
Derivative $E_0\omega \cos \omega t$ of electromotive force	Driving force $F_0 \cos \omega t$
Current $I(t)$	Displacement $y(t)$

Related to this analogy are **transducers**, devices that convert changes in a mechanical quantity (for instance, in a displacement) into changes in an electrical quantity that can be monitored; see Ref. [GenRef11] in App. 1.

PROBLEM SET 2.9

1–6 *RLC*-CIRCUITS: SPECIAL CASES

1. ***RC*-Circuit.** Model the *RC*-circuit in Fig. 64. Find the current due to a constant E.

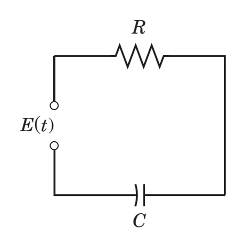


Fig. 64. *RC*-circuit

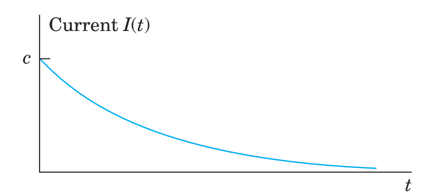


Fig. 65. Current 1 in Problem 1

2. ***RC*-Circuit.** Solve Prob. 1 when $E = E_0 \sin \omega t$ and R, C, E_0, and ω are arbitrary.
3. ***RL*-Circuit.** Model the *RL*-circuit in Fig. 66. Find a general solution when R, L, E are any constants. Graph or sketch solutions when $L = 0.25$ H, $R = 10\ \Omega$, and $E = 48$ V.

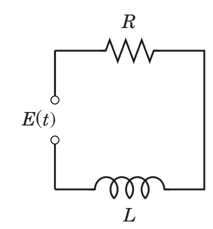


Fig. 66. *RL*-circuit

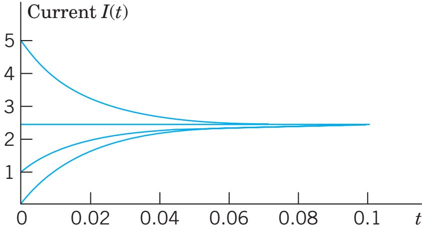


Fig. 67. Currents in Problem 3

4. ***RL*-Circuit.** Solve Prob. 3 when $E = E_0 \sin \omega t$ and R, L, E_0, and are arbitrary. Sketch a typical solution.

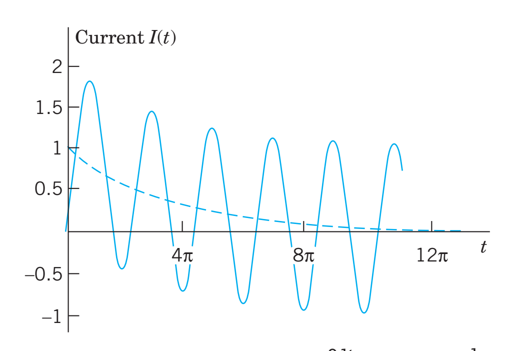


Fig. 68. Typical current $I = e^{-0.1t} + \sin(t - \frac{1}{4}\pi)$ in Problem 4

5. ***LC*-Circuit.** This is an *RLC*-circuit with negligibly small R (analog of an undamped mass–spring system). Find the current when $L = 0.5$ H, $C = 0.005$ F, and $E = \sin t$ V, assuming zero initial current and charge.

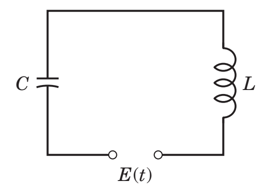


Fig. 69. *LC*-circuit

6. ***LC*-Circuit.** Find the current when $L = 0.5$ H, $C = 0.005$ F, $E = 2t^2$ V, and initial current and charge zero.

7–18 GENERAL *RLC*-CIRCUITS

7. **Tuning.** In tuning a stereo system to a radio station, we adjust the tuning control (turn a knob) that changes C (or perhaps L) in an *RLC*-circuit so that the amplitude of the steady-state current (5) becomes maximum. For what C will this happen?

8–14 Find the **steady-state current** in the *RLC*-circuit in Fig. 61 for the given data. Show the details of your work.

8. $R = 4\ \Omega, L = 0.5$ H, $C = 0.1$ F, $E = 500 \sin 2t$ V
9. $R = 4\ \Omega, L = 0.1$ H, $C = 0.05$ F, $E = 110$ V
10. $R = 2\ \Omega, L = 1$ H, $C = \frac{1}{20}$ F, $E = 157 \sin 3t$ V

11. $R = 12\ \Omega$, $L = 0.4$ H, $C = \frac{1}{80}$ F, $E = 220 \sin 10t$ V

12. $R = 0.2\ \Omega$, $L = 0.1$ H, $C = 2$ F, $E = 220 \sin 314t$ V

13. $R = 12$, $L = 1.2$ H, $C = \frac{20}{3} \cdot 10^{-3}$ F, $E = 12{,}000 \sin 25t$ V

14. Prove the claim in the text that if $R \neq 0$ (hence $R > 0$), then the transient current approaches I_p as $t \to \infty$.

15. Cases of damping. What are the conditions for an *RLC*-circuit to be (I) overdamped, (II) critically damped, (III) underdamped? What is the critical resistance R_{crit} (the analog of the critical damping constant $2\sqrt{mk}$)?

16–18 Solve the **initial value problem** for the *RLC*-circuit in Fig. 61 with the given data, assuming zero initial current and charge. Graph or sketch the solution. Show the details of your work.

16. $R = 8\ \Omega$, $L = 0.2$ H, $C = 12.5 \cdot 10^{-3}$ F, $E = 100 \sin 10t$ V

17. $R = 6\ \Omega$, $L = 1$ H, $C = 0.04$ F, $E = 600 (\cos t + 4 \sin t)$ V

18. $R = 18\ \Omega$, $L = 1$ H, $C = 12.5 \cdot 10^{-3}$ F, $E = 820 \cos 10t$ V

19. WRITING REPORT. Mechanic-Electric Analogy. Explain Table 2.2 in a 1–2 page report with examples, e.g., the analog (with $L = 1$ H) of a mass–spring system of mass 5 kg, damping constant 10 kg/sec, spring constant 60 kg/sec^2, and driving force $220 \cos 10t$ kg/sec.

20. Complex Solution Method. Solve $L\tilde{I}'' + R\tilde{I}' + \tilde{I}/C = E_0 e^{i\omega t}$, $i = \sqrt{-1}$, by substituting $I_p = Ke^{i\omega t}$ (K unknown) and its derivatives and taking the real part I_p of the solution $\tilde{I}_p$. Show agreement with (2), (4). *Hint*: Use (11) $e^{i\omega t} = \cos \omega t + i \sin \omega t$; cf. Sec. 2.2, and $i^2 = -1$.

2.10 Solution by Variation of Parameters

We continue our discussion of nonhomogeneous linear ODEs, that is

$$y'' + p(x)y' + q(x)y = r(x). \tag{1}$$

In Sec. 2.6 we have seen that a general solution of (1) is the sum of a general solution y_h of the corresponding homogeneous ODE and any particular solution y_p of (1). To obtain y_p when $r(x)$ is not too complicated, we can often use the *method of undetermined coefficients,* as we have shown in Sec. 2.7 and applied to basic engineering models in Secs. 2.8 and 2.9.

However, since this method is restricted to functions $r(x)$ whose derivatives are of a form similar to $r(x)$ itself (powers, exponential functions, etc.), it is desirable to have a method valid for more general ODEs (1), which we shall now develop. It is called the **method of variation of parameters** and is credited to Lagrange (Sec. 2.1). Here p, q, r in (1) may be variable (given functions of x), but we assume that they are continuous on some open interval I.

Lagrange's method gives a particular solution y_p of (1) on I in the form

$$y_p(x) = -y_1 \int \frac{y_2 r}{W}\, dx + y_2 \int \frac{y_1 r}{W}\, dx \tag{2}$$

where y_1, y_2 form a basis of solutions of the corresponding homogeneous ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{3}$$

on I, and W is the Wronskian of y_1, y_2,

$$W = y_1 y_2' - y_2 y_1' \qquad \text{(see Sec. 2.6).} \tag{4}$$

CAUTION! The solution formula (2) is obtained under the assumption that the ODE is written in standard form, with y'' as the first term as shown in (1). If it starts with $f(x)y''$, divide first by $f(x)$.

The integration in (2) may often cause difficulties, and so may the determination of y_1, y_2 if (1) has variable coefficients. If you have a choice, use the previous method. It is simpler. Before deriving (2) let us work an example for which you *do need* the new method. (Try otherwise.)

EXAMPLE 1 Method of Variation of Parameters

Solve the nonhomogeneous ODE

$$y'' + y = \sec x = \frac{1}{\cos x}.$$

Solution. A basis of solutions of the homogeneous ODE on any interval is $y_1 = \cos x$, $y_2 = \sin x$. This gives the Wronskian

$$W(y_1, y_2) = \cos x \cos x - \sin x\,(-\sin x) = 1.$$

From (2), choosing zero constants of integration, we get the particular solution of the given ODE

$$\begin{aligned} y_p &= -\cos x \int \sin x \sec x \, dx + \sin x \int \cos x \sec x \, dx \\ &= \cos x \ln |\cos x| + x \sin x \qquad \text{(Fig. 70)} \end{aligned}$$

Figure 70 shows y_p and its first term, which is small, so that $x \sin x$ essentially determines the shape of the curve of y_p. (Recall from Sec. 2.8 that we have seen $x \sin x$ in connection with resonance, except for notation.) From y_p and the general solution $y_h = c_1 y_1 + c_2 y_2$ of the homogeneous ODE we obtain the *answer*

$$y = y_h + y_p = (c_1 + \ln |\cos x|) \cos x + (c_2 + x) \sin x.$$

Had we included integration constants $-c_1, c_2$ in (2), then (2) would have given the additional $c_1 \cos x + c_2 \sin x = c_1 y_1 + c_2 y_2$, that is, a general solution of the given ODE directly from (2). This will always be the case. ■

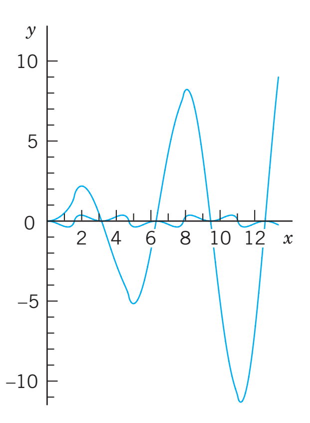


Fig. 70. Particular solution y_p and its first term in Example 1

Idea of the Method. Derivation of (2)

What idea did Lagrange have? What gave the method the name? Where do we use the continuity assumptions?

The idea is to start from a general solution

$$y_h(x) = c_1 y_1(x) + c_2 y_2(x)$$

of the homogeneous ODE (3) on an open interval I and to replace the constants ("the parameters") c_1 and c_2 by functions $u(x)$ and $v(x)$; this suggests the name of the method. We shall determine u and v so that the resulting function

$$y_p(x) = u(x)y_1(x) + v(x)y_2(x) \tag{5}$$

is a particular solution of the nonhomogeneous ODE (1). Note that y_h exists by Theorem 3 in Sec. 2.6 because of the continuity of p and q on I. (The continuity of r will be used later.)

We determine u and v by substituting (5) and its derivatives into (1). Differentiating (5), we obtain

$$y_p' = u'y_1 + uy_1' + v'y_2 + vy_2'.$$

Now y_p must satisfy (1). This is *one* condition for *two* functions u and v. It seems plausible that we may impose a *second* condition. Indeed, our calculation will show that we can determine u and v such that y_p satisfies (1) and u and v satisfy as a second condition the equation

$$u'y_1 + v'y_2 = 0. \tag{6}$$

This reduces the first derivative y_p' to the simpler form

$$y_p' = uy_1' + vy_2'. \tag{7}$$

Differentiating (7), we obtain

$$y_p'' = u'y_1' + uy_1'' + v'y_2' + vy_2''. \tag{8}$$

We now substitute y_p and its derivatives according to (5), (7), (8) into (1). Collecting terms in u and terms in v, we obtain

$$u(y_1'' + py_1' + qy_1) + v(y_2'' + py_2' + qy_2) + u'y_1' + v'y_2' = r.$$

Since y_1 and y_2 are solutions of the homogeneous ODE (3), this reduces to

$$u'y_1' + v'y_2' = r. \tag{9a}$$

Equation (6) is

$$u'y_1 + v'y_2 = 0. \tag{9b}$$

This is a linear system of two algebraic equations for the unknown functions u' and v'. We can solve it by elimination as follows (or by Cramer's rule in Sec. 7.6). To eliminate v', we multiply (9a) by $-y_2$ and (9b) by y_2' and add, obtaining

$$u'(y_1y_2' - y_2y_1') = -y_2r, \qquad \text{thus} \qquad u'W = -y_2r.$$

Here, W is the Wronskian (4) of y_1, y_2. To eliminate u' we multiply (9a) by y_1, and (9b) by $-y_1'$ and add, obtaining

$$v'(y_1y_2' - y_2y_1') = -y_1r, \qquad \text{thus} \qquad v'W = y_1r.$$

Since y_1, y_2 form a basis, we have $W \neq 0$ (by Theorem 2 in Sec. 2.6) and can divide by W,

$$u' = -\frac{y_2r}{W}, \qquad v' = \frac{y_1r}{W}. \tag{10}$$

By integration,

$$u = -\int \frac{y_2r}{W}\,dx, \qquad v = \int \frac{y_1r}{W}\,dx.$$

These integrals exist because $r(x)$ is continuous. Inserting them into (5) gives (2) and completes the derivation. ■

PROBLEM SET 2.10

1–13 GENERAL SOLUTION

Solve the given nonhomogeneous linear ODE by variation of parameters or undetermined coefficients. Show the details of your work.

1. $y'' + 9y = \sec 3x$
2. $y'' + 9y = \csc 3x$
3. $x^2y'' - 2xy' + 2y = x^3 \sin x$
4. $y'' - 4y' + 5y = e^{2x} \csc x$
5. $y'' + y = \cos x - \sin x$
6. $(D^2 + 6D + 9I)y = 16e^{-3x}/(x^2 + 1)$
7. $(D^2 - 4D + 4I)y = 6e^{2x}/x^4$
8. $(D^2 + 4I)y = \cosh 2x$
9. $(D^2 - 2D + I)y = 35x^{3/2}e^x$
10. $(D^2 + 2D + 2I)y = 4e^{-x} \sec^3 x$
11. $(x^2D^2 - 4xD + 6I)y = 21x^{-4}$
12. $(D^2 - I)y = 1/\cosh x$
13. $(x^2D^2 + xD - 9I)y = 48x^5$
14. **TEAM PROJECT. Comparison of Methods. Invention.** The undetermined-coefficient method should be used whenever possible because it is simpler. Compare it with the present method as follows.

    (a) Solve $y'' + 4y' + 3y = 65 \cos 2x$ by both methods, showing all details, and compare.

    (b) Solve $y'' - 2y' + y = r_1 + r_2, r_1 = 35x^{3/2}e^x r_2 = x^2$ by applying each method to a suitable function on the right.

    (c) Experiment to invent an undetermined-coefficient method for nonhomogeneous Euler–Cauchy equations.

CHAPTER 2 REVIEW QUESTIONS AND PROBLEMS

1. Why are linear ODEs preferable to nonlinear ones in modeling?
2. What does an initial value problem of a second-order ODE look like? Why must you have a general solution to solve it?
3. By what methods can you get a general solution of a nonhomogeneous ODE from a general solution of a homogeneous one?
4. Describe applications of ODEs in mechanical systems. What are the electrical analogs of the latter?
5. What is resonance? How can you remove undesirable resonance of a construction, such as a bridge, a ship, or a machine?
6. What do you know about existence and uniqueness of solutions of linear second-order ODEs?

7–18 GENERAL SOLUTION

Find a general solution. Show the details of your calculation.

7. $4y'' + 32y' + 63y = 0$
8. $y'' + y' - 12y = 0$
9. $y'' + 6y' + 34y = 0$
10. $y'' + 0.20y' + 0.17y = 0$
11. $(100D^2 - 160D + 64I)y = 0$
12. $(D^2 + 4\pi D + 4\pi^2I)y = 0$
13. $(x^2D^2 + 2xD - 12I)y = 0$
14. $(x^2D^2 + xD - 9I)y = 0$
15. $(2D^2 - 3D - 2I)y = 13 - 2x^2$
16. $(D^2 + 2D + 2I)y = 3e^{-x} \cos 2x$
17. $(4D^2 - 12D + 9I)y = 2e^{1.5x}$
18. $yy'' = 2y'^2$

19–22 INITIAL VALUE PROBLEMS

Solve the problem, showing the details of your work. Sketch or graph the solution.

19. $y'' + 16y = 17e^x, \quad y(0) = 6, \quad y'(0) = -2$

20. $y'' - 3y' + 2y = 10 \sin x, \quad y(0) = 1, \quad y'(0) = -6$

21. $(x^2D^2 + xD - I)y = 16x^3, \quad y(1) = -1, \quad y'(1) = 1$

22. $(x^2D^2 + 15xD + 49I)y = 0, \quad y(1) = 2,$ $y'(1) = -11$

23–30 APPLICATIONS

23. Find the steady-state current in the *RLC*-circuit in Fig. 71 when $R = 2\text{ k}\Omega\ (2000\ \Omega)$, $L = 1$ H, $C = 4 \cdot 10^{-3}$ F, and $E = 110 \sin 415t$ V (66 cycles/sec).

24. Find a general solution of the homogeneous linear ODE corresponding to the ODE in Prob. 23.

25. Find the steady-state current in the *RLC*-circuit in Fig. 71 when $R = 50\ \Omega$, $L = 30$ H, $C = 0.025$ F, $E = 200 \sin 4t$ V.

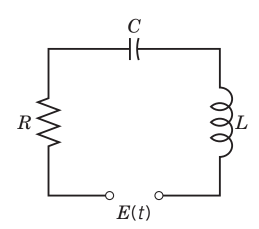


Fig. 71. *RLC*-circuit

26. Find the current in the *RLC*-circuit in Fig. 71 when $R = 40\ \Omega$, $L = 0.4$ H, $C = 10^{-4}$ F, $E = 220 \sin 314t$ V (50 cycles/sec).

27. Find an electrical analog of the mass–spring system with mass 4 kg, spring constant 10 kg/sec^2, damping constant 20 kg/sec, and driving force $100 \sin 4t$ nt.

28. Find the motion of the mass–spring system in Fig. 72 with mass 0.125 kg, damping 0, spring constant 1.125 kg/sec^2, and driving force $\cos t - 4 \sin t$ nt, assuming zero initial displacement and velocity. For what frequency of the driving force would you get resonance?

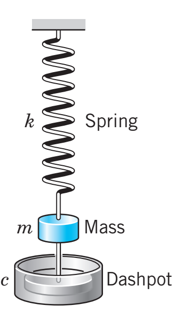


Fig. 72. Mass–spring system

29. Show that the system in Fig. 72 with $m = 4$, $c = 0$, $k = 36$, and driving force $61 \cos 3.1t$ exhibits beats. *Hint:* Choose zero initial conditions.

30. In Fig. 72, let $m = 1$ kg, $c = 4$ kg/sec, $k = 24$ kg/sec^2, and $r(t) = 10 \cos \omega t$ nt. Determine w such that you get the steady-state vibration of maximum possible amplitude. Determine this amplitude. Then find the general solution with this ω and check whether the results are in agreement.

SUMMARY OF CHAPTER 2
Second-Order Linear ODEs

Second-order linear ODEs are particularly important in applications, for instance, in mechanics (Secs. 2.4, 2.8) and electrical engineering (Sec. 2.9). A second-order ODE is called **linear** if it can be written

$$(1) \qquad y'' + p(x)y' + q(x)y = r(x) \qquad \text{(Sec. 2.1).}$$

(If the first term is, say, $f(x)y''$, divide by $f(x)$ to get the "**standard form**" (1) with y'' as the first term.) Equation (1) is called **homogeneous** if $r(x)$ is zero for all x considered, usually in some open interval; this is written $r(x) \equiv 0$. Then

$$(2) \qquad y'' + p(x)y' + q(x)y = 0.$$

Equation (1) is called **nonhomogeneous** if $r(x) \not\equiv 0$ (meaning $r(x)$ is not zero for some x considered).

For the homogeneous ODE (2) we have the important **superposition principle** (Sec. 2.1) that a linear combination $y = ky_1 + ly_2$ of two solutions y_1, y_2 is again a solution.

Two *linearly independent* solutions y_1, y_2 of (2) on an open interval I form a **basis** (or **fundamental system**) of solutions on I. and $y = c_1y_1 + c_2y_2$ with arbitrary constants c_1, c_2 a **general solution** of (2) on I. From it we obtain a **particular solution** if we specify numeric values (numbers) for c_1 and c_2, usually by prescribing two **initial conditions**

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \qquad (x_0, K_0, K_1 \text{ given numbers; Sec. 2.1}). \tag{3}$$

(2) and (3) together form an **initial value problem**. Similarly for (1) and (3).

For a nonhomogeneous ODE (1) a **general solution** is of the form

$$y = y_h + y_p \qquad \text{(Sec. 2.7)}. \tag{4}$$

Here y_h is a general solution of (2) and y_p is a particular solution of (1). Such a y_p can be determined by a general method (***variation of parameters***, Sec. 2.10) or in many practical cases by the ***method of undetermined coefficients***. The latter applies when (1) has constant coefficients p and q, and $r(x)$ is a power of x, sine, cosine, etc. (Sec. 2.7). Then we write (1) as

$$y'' + ay' + by = r(x) \qquad \text{(Sec. 2.7)}. \tag{5}$$

The corresponding homogeneous ODE $y' + ay' + by = 0$ has solutions $y = e^{\lambda x}$, where λ is a root of

$$\lambda^2 + a\lambda + b = 0. \tag{6}$$

Hence there are three cases (Sec. 2.2):

Case	Type of Roots	General Solution
I	Distinct real λ_1, λ_2	$y = c_1e^{\lambda_1 x} + c_2e^{\lambda_2 x}$
II	Double $-\frac{1}{2}a$	$y = (c_1 + c_2x)e^{-ax/2}$
III	Complex $-\frac{1}{2}a \pm i\omega^*$	$y = e^{-ax/2}(A \cos \omega^* x + B \sin \omega^* x)$

Here ω^* is used since ω is needed in driving forces.

Important applications of (5) in mechanical and electrical engineering in connection with ***vibrations*** and ***resonance*** are discussed in Secs. 2.4, 2.7, and 2.8.

Another large class of ODEs solvable "algebraically" consists of the **Euler–Cauchy equations**

$$x^2y'' + axy' + by = 0 \qquad \text{(Sec. 2.5)}. \tag{7}$$

These have solutions of the form $y = x^m$, where m is a solution of the auxiliary equation

$$m^2 + (a - 1)m + b = 0. \tag{8}$$

Existence and uniqueness of solutions of (1) and (2) is discussed in Secs. 2.6 and 2.7, and ***reduction of order*** in Sec. 2.1.

CHAPTER 3

Higher Order Linear ODEs

The concepts and methods of solving linear ODEs of order $n = 2$ extend nicely to linear ODEs of higher order n, that is, $n = 3, 4$, etc. This shows that the theory explained in Chap. 2 for second-order linear ODEs is attractive, since it can be extended in a straightforward way to arbitrary n. We do so in this chapter and notice that the formulas become more involved, the variety of roots of the characteristic equation (in Sec. 3.2) becomes much larger with increasing n, and the Wronskian plays a more prominent role.

The concepts and methods of solving second-order linear ODEs extend readily to linear ODEs of higher order.

This chapter follows Chap. 2 naturally, since the results of Chap. 2 can be readily extended to that of Chap. 3.

Prerequisite: Secs. 2.1, 2.2, 2.6, 2.7, 2.10.
References and Answers to Problems: App. 1 Part A, and App. 2.

3.1 Homogeneous Linear ODEs

Recall from Sec. 1.1 that an ODE is of ***n*th order** if the nth derivative $y^{(n)} = d^n y/dx^n$ of the unknown function $y(x)$ is the highest occurring derivative. Thus the ODE is of the form

$$F(x, y, y', \cdots, y^{(n)}) = 0$$

where lower order derivatives and y itself may or may not occur. Such an ODE is called **linear** if it can be written

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x). \tag{1}$$

(For $n = 2$ this is (1) in Sec. 2.1 with $p_1 = p$ and $p_0 = q$.) The **coefficients** $p_0, \cdots, p_{n-1}$ and the function r on the right are any given functions of x, and y is unknown. $y^{(n)}$ has coefficient 1. We call this the **standard form**. (If you have $p_n(x)y^{(n)}$, divide by $p_n(x)$ to get this form.) An nth-order ODE that cannot be written in the form (1) is called **nonlinear**.

If $r(x)$ is identically zero, $r(x) \equiv 0$ (zero for all x considered, usually in some open interval I), then (1) becomes

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0 \tag{2}$$

and is called **homogeneous**. If $r(x)$ is not identically zero, then the ODE is called **nonhomogeneous**. This is as in Sec. 2.1.

A **solution** of an nth-order (linear or nonlinear) ODE on some open interval I is a function $y = h(x)$ that is defined and n times differentiable on I and is such that the ODE becomes an identity if we replace the unknown function y and its derivatives by h and its corresponding derivatives.

Sections 3.1–3.2 will be devoted to homogeneous linear ODEs and Section 3.3 to nonhomogeneous linear ODEs.

Homogeneous Linear ODE: Superposition Principle, General Solution

The basic **superposition or linearity principle** of Sec. 2.1 extends to nth order homogeneous linear ODEs as follows.

THEOREM 1

Fundamental Theorem for the Homogeneous Linear ODE (2)

For a homogeneous linear ODE (2), *sums and constant multiples of solutions on some open interval I are again solutions on I.* (This does ***not*** hold for a nonhomogeneous or nonlinear ODE!)

The proof is a simple generalization of that in Sec. 2.1 and we leave it to the student.

Our further discussion parallels and extends that for second-order ODEs in Sec. 2.1. So we next define a general solution of (2), which will require an extension of linear independence from 2 to n functions.

DEFINITION

General Solution, Basis, Particular Solution

A **general solution** of (2) on an open interval I is a solution of (2) on I of the form

$$y(x) = c_1 y_1(x) + \cdots + c_n y_n(x) \qquad (c_1, \cdots, c_n \text{ arbitrary}) \tag{3}$$

where $y_1, \cdots, y_n$ is a **basis** (or **fundamental system**) of solutions of (2) on I; that is, these solutions are linearly independent on I, as defined below.

A **particular solution** of (2) on I is obtained if we assign specific values to the n constants $c_1, \cdots, c_n$ in (3).

DEFINITION

Linear Independence and Dependence

Consider n functions $y_1(x), \cdots, y_n(x)$ defined *on some interval I.*

These functions are called **linearly independent** on I if the equation

$$k_1 y_1(x) + \cdots + k_n y_n(x) = 0 \qquad \text{on } I \tag{4}$$

implies that all $k_1, \cdots, k_n$ are zero. These functions are called **linearly dependent** on I if this equation also holds on I for some $k_1, \cdots, k_n$ not all zero.

If and only if $y_1, \cdots, y_n$ are linearly dependent on I, we can express (at least) one of these functions on I as a "**linear combination**" of the other $n - 1$ functions, that is, as a sum of those functions, each multiplied by a constant (zero or not). This motivates the term "linearly dependent." For instance, if (4) holds with $k_1 \neq 0$, we can divide by k_1 and express y_1 as the linear combination

$$y_1 = -\frac{1}{k_1}(k_2 y_2 + \cdots + k_n y_n).$$

Note that when $n = 2$, these concepts reduce to those defined in Sec. 2.1.

EXAMPLE 1 **Linear Dependence**

Show that the functions $y_1 = x^2$, $y_2 = 5x$, $y_3 = 2x$ are linearly dependent on any interval.

Solution. $y_2 = 0y_1 + 2.5y_3$. This proves linear dependence on any interval. ■

EXAMPLE 2 **Linear Independence**

Show that $y_1 = x$, $y_2 = x^2$, $y_3 = x^3$ are linearly independent on any interval, for instance, on $-1 \leqq x \leqq 2$.

Solution. Equation (4) is $k_1 x + k_2 x^2 + k_3 x^3 = 0$. Taking (a) $x = -1$, (b) $x = 1$, (c) $x = 2$, we get

$$\text{(a) } -k_1 + k_2 - k_3 = 0, \qquad \text{(b) } k_1 + k_2 + k_3 = 0, \qquad \text{(c) } 2k_1 + 4k_2 + 8k_3 = 0.$$

$k_2 = 0$ from (a) + (b). Then $k_3 = 0$ from (c) -2(b). Then $k_1 = 0$ from (b). This proves linear independence.

A better method for testing linear independence of solutions of ODEs will soon be explained. ■

EXAMPLE 3 **General Solution. Basis**

Solve the fourth-order ODE

$$y^{\text{iv}} - 5y'' + 4y = 0 \qquad \text{(where } y^{\text{iv}} = d^4y/dx^4\text{).}$$

Solution. As in Sec. 2.2 we substitute $y = e^{\lambda x}$. Omitting the common factor $e^{\lambda x}$, we obtain the characteristic equation

$$\lambda^4 - 5\lambda^2 + 4 = 0.$$

This is a quadratic equation in $\mu = \lambda^2$, namely,

$$\mu^2 - 5\mu + 4 = (\mu - 1)(\mu - 4) = 0.$$

The roots are $\mu = 1$ and 4. Hence $\lambda = -2, -1, 1, 2$. This gives four solutions. A general solution on any interval is

$$y = c_1 e^{-2x} + c_2 e^{-x} + c_3 e^{x} + c_4 e^{2x}$$

provided those four solutions are linearly independent. This is true but will be shown later. ■

Initial Value Problem. Existence and Uniqueness

An **initial value problem** for the ODE (2) consists of (2) and n **initial conditions**

$$\text{(5)} \qquad y(x_0) = K_0, \qquad y'(x_0) = K_1, \qquad \cdots, \qquad y^{(n-1)}(x_0) = K_{n-1}$$

with given x_0 in the open interval I considered, and given $K_0, \cdots, K_{n-1}$.

In extension of the existence and uniqueness theorem in Sec. 2.6 we now have the following.

THEOREM 2

Existence and Uniqueness Theorem for Initial Value Problems

If the coefficients $p_0(x), \cdots, p_{n-1}(x)$ *of* (2) *are continuous on some open interval* I *and* x_0 *is in* I, *then the initial value problem* (2), (5) *has a unique solution* $y(x)$ *on* I.

Existence is proved in Ref. [A11] in App. 1. Uniqueness can be proved by a slight generalization of the uniqueness proof at the beginning of App. 4.

EXAMPLE 4

Initial Value Problem for a Third-Order Euler–Cauchy Equation

Solve the following initial value problem on any open interval I on the positive x-axis containing $x = 1$.

$$x^3y''' - 3x^2y'' + 6xy' - 6y = 0, \qquad y(1) = 2, \qquad y'(1) = 1, \qquad y''(1) = -4.$$

Solution. ***Step 1. General solution.*** As in Sec. 2.5 we try $y = x^m$. By differentiation and substitution,

$$m(m-1)(m-2)x^m - 3m(m-1)x^m + 6mx^m - 6x^m = 0.$$

Dropping x^m and ordering gives $m^3 - 6m^2 + 11m - 6 = 0$. If we can guess the root $m = 1$. We can divide by $m - 1$ and find the other roots 2 and 3, thus obtaining the solutions x, x^2, x^3, which are linearly independent on I (see Example 2). [In general one shall need a root-finding method, such as Newton's (Sec. 19.2), also available in a CAS (Computer Algebra System).] Hence a general solution is

$$y = c_1x + c_2x^2 + c_3x^3$$

valid on any interval I, even when it includes $x = 0$ where the coefficients of the ODE divided by x^3 (to have the standard form) are not continuous.

Step 2. Particular solution. The derivatives are $y' = c_1 + 2c_2x + 3c_3x^2$ and $y'' = 2c_2 + 6c_3x$. From this, and y and the initial conditions, we get by setting $x = 1$

$$\begin{aligned}
&\text{(a)}\ \ y(1) = c_1 + c_2 + c_3 = 2\\
&\text{(b)}\ \ y'(1) = c_1 + 2c_2 + 3c_3 = 1\\
&\text{(c)}\ \ y''(1) = 2c_2 + 6c_3 = -4.
\end{aligned}$$

This is solved by Cramer's rule (Sec. 7.6), or by elimination, which is simple, as follows. (b) − (a) gives (d) $c_2 + 2c_3 = -1$. Then (c) − 2(d) gives $c_3 = -1$. Then (c) gives $c_2 = 1$. Finally $c_1 = 2$ from (a).
Answer: $y = 2x + x^2 - x^3$. ■

Linear Independence of Solutions. Wronskian

Linear independence of solutions is crucial for obtaining general solutions. Although it can often be seen by inspection, it would be good to have a criterion for it. Now Theorem 2 in Sec. 2.6 extends from order $n = 2$ to any n. This extended criterion uses the **Wronskian** W of n solutions $y_1, \cdots, y_n$ defined as the nth-order determinant

$$W(y_1, \cdots, y_n) = \begin{vmatrix} y_1 & y_2 & \cdots & y_n \\ y_1' & y_2' & \cdots & y_n' \\ \cdot & \cdot & \cdots & \cdot \\ y_1^{(n-1)} & y_2^{(n-1)} & \cdots & y_n^{(n-1)} \end{vmatrix}. \tag{6}$$

Note that W depends on x since $y_1, \cdots, y_n$ do. The criterion states that these solutions form a basis if and only if W is not zero; more precisely:

THEOREM 3

Linear Dependence and Independence of Solutions

Let the ODE (2) have continuous coefficients $p_0(x), \cdots, p_{n-1}(x)$ on an open interval I. Then n solutions $y_1, \cdots, y_n$ of (2) on I are linearly dependent on I if and only if their Wronskian is zero for some $x = x_0$ in I. Furthermore, if W is zero for $x = x_0$, then W is identically zero on I. Hence if there is an x_1 in I at which W is not zero, then $y_1, \cdots, y_n$ are linearly independent on I, so that they form a basis of solutions of (2) on I.

PROOF **(a)** Let $y_1, \cdots, y_n$ be linearly dependent solutions of (2) on I. Then, by definition, there are constants $k_1, \cdots, k_n$ *not all zero*, such that for all x in I,

$$k_1 y_1 + \cdots + k_n y_n = 0. \tag{7}$$

By $n - 1$ differentiations of (7) we obtain for all x in I

$$\begin{aligned} k_1 y_1' + \cdots + k_n y_n' &= 0 \\ &\vdots \\ k_1 y_1^{(n-1)} + \cdots + k_n y_n^{(n-1)} &= 0. \end{aligned} \tag{8}$$

(7), (8) is a homogeneous linear system of algebraic equations with a nontrivial solution $k_1, \cdots, k_n$. Hence its coefficient determinant must be zero for every x on I, by Cramer's theorem (Sec. 7.7). But that determinant is the Wronskian W, as we see from (6). Hence W is zero for every x on I.

(b) Conversely, if W is zero at an x_0 in I, then the system (7), (8) with $x = x_0$ has a solution $k_1^*, \cdots, k_n^*$, not all zero, by the same theorem. With these constants we define the solution $y^* = k_1^* y_1 + \cdots + k_n^* y_n$ of (2) on I. By (7), (8) this solution satisfies the initial conditions $y^*(x_0) = 0, \cdots, y^{*(n-1)}(x_0) = 0$. But another solution satisfying the same conditions is $y \equiv 0$. Hence $y^* \equiv y$ by Theorem 2, which applies since the coefficients of (2) are continuous. Together, $y^* = k_1^* y_1 + \cdots + k_n^* y_n \equiv 0$ on I. This means linear dependence of $y_1, \cdots, y_n$ on I.

(c) If W is zero at an x_0 in I, we have linear dependence by (b) and then $W \equiv 0$ by (a). Hence if W is not zero at an x_1 in I, the solutions $y_1, \cdots, y_n$ must be linearly independent on I. ■

EXAMPLE 5 **Basis, Wronskian**

We can now prove that in Example 3 we do have a basis. In evaluating W, pull out the exponential functions columnwise. In the result, subtract Column 1 from Columns 2, 3, 4 (without changing Column 1). Then expand by Row 1. In the resulting third-order determinant, subtract Column 1 from Column 2 and expand the result by Row 2:

$$W = \begin{vmatrix} e^{-2x} & e^{-x} & e^{x} & e^{2x} \\ -2e^{-2x} & -e^{-x} & e^{x} & 2e^{2x} \\ 4e^{-2x} & e^{-x} & e^{x} & 4e^{2x} \\ -8e^{-2x} & -e^{-x} & e^{x} & 8e^{2x} \end{vmatrix} = \begin{vmatrix} 1 & 1 & 1 & 1 \\ -2 & -1 & 1 & 2 \\ 4 & 1 & 1 & 4 \\ -8 & -1 & 1 & 8 \end{vmatrix} = \begin{vmatrix} 1 & 3 & 4 \\ -3 & -3 & 0 \\ 7 & 9 & 16 \end{vmatrix} = 72.$$ ■

A General Solution of (2) Includes All Solutions

Let us first show that general solutions always exist. Indeed, Theorem 3 in Sec. 2.6 extends as follows.

THEOREM 4

Existence of a General Solution

If the coefficients $p_0(x), \cdots, p_{n-1}(x)$ of (2) are continuous on some open interval I, then (2) has a general solution on I.

PROOF We choose any fixed x_0 in I. By Theorem 2 the ODE (2) has n solutions $y_1, \cdots, y_n$, where y_j satisfies initial conditions (5) with $K_{j-1} = 1$ and all other K's equal to zero. Their Wronskian at x_0 equals 1. For instance, when $n = 3$, then $y_1(x_0) = 1, y_2'(x_0) = 1$, $y_3''(x_0) = 1$, and the other initial values are zero. Thus, as claimed,

$$W(y_1(x_0), y_2(x_0), y_3(x_0)) = \begin{vmatrix} y_1(x_0) & y_2(x_0) & y_3(x_0) \\ y_1'(x_0) & y_2'(x_0) & y_3'(x_0) \\ y_1''(x_0) & y_2''(x_0) & y_3''(x_0) \end{vmatrix} = \begin{vmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{vmatrix} = 1.$$

Hence for any n those solutions $y_1, \cdots, y_n$ are linearly independent on I, by Theorem 3. They form a basis on I, and $y = c_1 y_1 + \cdots + c_n y_n$ is a general solution of (2) on I. ■

We can now prove the basic property that, from a general solution of (2), every solution of (2) can be obtained by choosing suitable values of the arbitrary constants. Hence an nth-order ***linear*** ODE has no **singular solutions**, that is, solutions that cannot be obtained from a general solution.

THEOREM 5

General Solution Includes All Solutions

If the ODE (2) has continuous coefficients $p_0(x), \cdots, p_{n-1}(x)$ on some open interval I, then every solution $y = Y(x)$ of (2) on I is of the form

$$Y(x) = C_1 y_1(x) + \cdots + C_n y_n(x) \tag{9}$$

where $y_1, \cdots, y_n$ is a basis of solutions of (2) on I and $C_1, \cdots, C_n$ are suitable constants.

PROOF Let Y be a given solution and $y = c_1 y_1 + \cdots + c_n y_n$ a general solution of (2) on I. We choose any fixed x_0 in I and show that we can find constants $c_1, \cdots, c_n$ for which y and its first $n - 1$ derivatives agree with Y and its corresponding derivatives at x_0. That is, we should have at $x = x_0$

$$\begin{aligned} c_1 y_1 + \cdots + \quad c_n y_n &= Y \\ c_1 y_1' + \cdots + \quad c_n y_n' &= Y' \\ &\vdots \\ c_1 y_1^{(n-1)} + \cdots + c_n y_n^{(n-1)} &= Y^{(n-1)}. \end{aligned} \tag{10}$$

But this is a linear system of equations in the unknowns $c_1, \cdots, c_n$. Its coefficient determinant is the Wronskian W of $y_1, \cdots, y_n$ at x_0. Since $y_1, \cdots, y_n$ form a basis, they

are linearly independent, so that W is not zero by Theorem 3. Hence (10) has a unique solution $c_1 = C_1, \cdots, c_n = C_n$ (by Cramer's theorem in Sec. 7.7). With these values we obtain the particular solution

$$y^*(x) = C_1 y_1(x) + \cdots + C_n y_n(x)$$

on I. Equation (10) shows that y^* and its first $n - 1$ derivatives agree at x_0 with Y and its corresponding derivatives. That is, y^* and Y satisfy, at x_0, the same initial conditions. The uniqueness theorem (Theorem 2) now implies that $y^* \equiv Y$ on I. This proves the theorem. ■

This completes our theory of the homogeneous linear ODE (2). Note that for $n = 2$ it is identical with that in Sec. 2.6. This had to be expected.

PROBLEM SET 3.1

1–6 BASES: TYPICAL EXAMPLES

To get a feel for higher order ODEs, show that the given functions are solutions and form a basis on any interval. Use Wronskians. In Prob. 6, $x > 0$,

1. $1, x, x^2, x^3, \quad y^{iv} = 0$
2. $e^x, e^{-x}, e^{2x}, \quad y''' - 2y'' - y' + 2y = 0$
3. $\cos x, \sin x, x \cos x, x \sin x, \quad y^{iv} + 2y'' + y = 0$
4. $e^{-4x}, xe^{-4x}, x^2e^{-4x}, \quad y''' + 12y'' + 48y' + 64y = 0$
5. $1, e^{-x} \cos 2x, e^{-x} \sin 2x, \quad y''' + 2y'' + 5y' = 0$
6. $1, x^2, x^4, \quad x^2y''' - 3xy'' + 3y' = 0$
7. **TEAM PROJECT. General Properties of Solutions of Linear ODEs.** These properties are important in obtaining new solutions from given ones. Therefore extend Team Project 38 in Sec. 2.2 to nth-order ODEs. Explore statements on sums and multiples of solutions of (1) and (2) systematically and with proofs. Recognize clearly that no new ideas are needed in this extension from $n = 2$ to general n.

8–15 LINEAR INDEPENDENCE

Are the given functions linearly independent or dependent on the half-axis $x \geq 0$? Give reason.

8. $x^2, 1/x^2, 0$
9. $\tan x, \cot x, 1$
10. $e^{2x}, xe^{2x}, x^2e^{2x}$
11. $e^x \cos x, e^x \sin x, e^x$
12. $\sin^2 x, \cos^2 x, \cos 2x$
13. $\sin x, \cos x, \sin 2x$
14. $\cos^2 x, \sin^2 x, 2\pi$
15. $\cosh 2x, \sinh 2x, e^{2x}$
16. **TEAM PROJECT. Linear Independence and Dependence. (a)** Investigate the given question about a set S of functions on an interval I. Give an example. Prove your answer.

(1) If S contains the zero function, can S be linearly independent?

(2) If S is linearly independent on a subinterval J of I, is it linearly independent on I?

(3) If S is linearly dependent on a subinterval J of I, is it linearly dependent on I?

(4) If S is linearly independent on I, is it linearly independent on a subinterval J?

(5) If S is linearly dependent on I, is it linearly independent on a subinterval J?

(6) If S is linearly dependent on I, and if T contains S, is T linearly dependent on I?

(b) In what cases can you use the Wronskian for testing linear independence? By what other means can you perform such a test?

3.2 Homogeneous Linear ODEs with Constant Coefficients

We proceed along the lines of Sec. 2.2, and generalize the results from $n = 2$ to arbitrary n. We want to solve an nth-order homogeneous linear ODE with constant coefficients, written as

$$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1y' + a_0y = 0 \qquad (1)$$

where $y^{(n)} = d^n y/dx^n$, etc. As in Sec. 2.2, we substitute $y = e^{\lambda x}$ to obtain the characteristic equation

$$\lambda^{(n)} + a_{n-1}\lambda^{(n-1)} + \cdots + a_1\lambda + a_0 y = 0 \tag{2}$$

of (1). If λ is a root of (2), then $y = e^{\lambda x}$ is a solution of (1). To find these roots, you may need a numeric method, such as Newton's in Sec. 19.2, also available on the usual CASs. For general n there are more cases than for $n = 2$. We can have distinct real roots, simple complex roots, multiple roots, and multiple complex roots, respectively. This will be shown next and illustrated by examples.

Distinct Real Roots

If all the n roots $\lambda_1, \cdots, \lambda_n$ of (2) are real and different, then the n solutions

$$y_1 = e^{\lambda_1 x}, \qquad \cdots, \qquad y_n = e^{\lambda_n x}. \tag{3}$$

constitute a basis for all x. The corresponding general solution of (1) is

$$y = c_1 e^{\lambda_1 x} + \cdots + c_n e^{\lambda_n x}. \tag{4}$$

Indeed, the solutions in (3) are linearly independent, as we shall see after the example.

EXAMPLE 1 Distinct Real Roots

Solve the ODE $y''' - 2y'' - y' + 2y = 0$.

Solution. The characteristic equation is $\lambda^3 - 2\lambda^2 - \lambda + 2 = 0$. It has the roots $-1, 1, 2$; if you find one of them by inspection, you can obtain the other two roots by solving a quadratic equation (explain!). The corresponding general solution (4) is $y = c_1 e^{-x} + c_2 e^{x} + c_3 e^{2x}$. ■

Linear Independence of (3). Students familiar with nth-order determinants may verify that, by pulling out all exponential functions from the columns and denoting their product by $E = \exp[(\lambda_1 + \cdots + \lambda_n)x]$, the Wronskian of the solutions in (3) becomes

$$W = \begin{vmatrix} e^{\lambda_1 x} & e^{\lambda_2 x} & \cdots & e^{\lambda_n x} \\ \lambda_1 e^{\lambda_1 x} & \lambda_2 e^{\lambda_2 x} & \cdots & \lambda_n e^{\lambda_n x} \\ \lambda_1^2 e^{\lambda_1 x} & \lambda_2^2 e^{\lambda_2 x} & \cdots & \lambda_n^2 e^{\lambda_n x} \\ \cdot & \cdot & \cdots & \cdot \\ \lambda_1^{n-1} e^{\lambda_1 x} & \lambda_2^{n-1} e^{\lambda_2 x} & \cdots & \lambda_n^{n-1} e^{\lambda_n x} \end{vmatrix}$$

$$= E \begin{vmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \cdot & \cdot & \cdots & \cdot \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{vmatrix}. \tag{5}$$

The exponential function E is never zero. Hence $W = 0$ if and only if the determinant on the right is zero. This is a so-called **Vandermonde** or **Cauchy determinant**.[1] It can be shown that it equals

$$(-1)^{n(n-1)/2}V \tag{6}$$

where V is the product of all factors $\lambda_j - \lambda_k$ with $j < k\ (\leqq n)$; for instance, when $n = 3$ we get $-V = -(\lambda_1 - \lambda_2)(\lambda_1 - \lambda_3)(\lambda_2 - \lambda_3)$. This shows that the Wronskian is not zero if and only if all the n roots of (2) are different and thus gives the following.

THEOREM 1

Basis

Solutions $y_1 = e^{\lambda_1 x}, \cdots, y_n = e^{\lambda_n x}$ *of* (1) (*with any real or complex* λ_j*'s*) *form a basis of solutions of* (1) *on any open interval if and only if all n roots of* (2) *are different.*

Actually, Theorem 1 is an important special case of our more general result obtained from (5) and (6):

THEOREM 2

Linear Independence

Any number of solutions of (1) *of the form* $e^{\lambda x}$ *are linearly independent on an open interval I if and only if the corresponding* λ *are all different.*

Simple Complex Roots

If complex roots occur, they must occur in conjugate pairs since the coefficients of (1) are real. Thus, if $\lambda = \gamma + i\omega$ is a simple root of (2), so is the conjugate $\bar{\lambda} = \gamma - i\omega$, and two corresponding linearly independent solutions are (as in Sec. 2.2, except for notation)

$$y_1 = e^{\gamma x}\cos \omega x, \qquad y_2 = e^{\gamma x}\sin \omega x.$$

EXAMPLE 2 **Simple Complex Roots. Initial Value Problem**

Solve the initial value problem

$$y''' - y'' + 100y' - 100y = 0, \qquad y(0) = 4, \qquad y'(0) = 11, \qquad y''(0) = -299.$$

Solution. The characteristic equation is $\lambda^3 - \lambda^2 + 100\lambda - 100 = 0$. It has the root 1, as can perhaps be seen by inspection. Then division by $\lambda - 1$ shows that the other roots are $\pm 10i$. Hence a general solution and its derivatives (obtained by differentiation) are

$$y = c_1 e^x + A\cos 10x + B\sin 10x,$$

$$y' = c_1 e^x - 10A\sin 10x + 10B\cos 10x,$$

$$y'' = c_1 e^x - 100A\cos 10x - 100B\sin 10x.$$

---

[1]ALEXANDRE THÉOPHILE VANDERMONDE (1735–1796), French mathematician, who worked on solution of equations by determinants. For CAUCHY see footnote 4, in Sec. 2.5.

From this and the initial conditions we obtain, by setting $x = 0$,

$$\text{(a)}\ c_1 + A = 4, \qquad \text{(b)}\ c_1 + 10B = 11, \qquad \text{(c)}\ c_1 - 100A = -299.$$

We solve this system for the unknowns A, B, c_1. Equation (a) minus Equation (c) gives $101A = 303$, $A = 3$. Then $c_1 = 1$ from (a) and $B = 1$ from (b). The solution is (Fig. 73)

$$y = e^x + 3 \cos 10x + \sin 10x.$$

This gives the solution curve, which oscillates about e^x (dashed in Fig. 73). ■

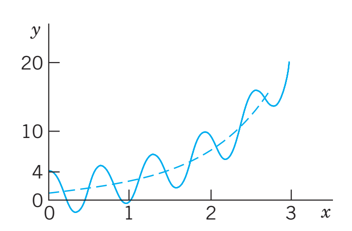


Fig. 73. Solution in Example 2

Multiple Real Roots

If a real double root occurs, say, $\lambda_1 = \lambda_2$, then $y_1 = y_2$ in (3), and we take y_1 and xy_1 as corresponding linearly independent solutions. This is as in Sec. 2.2.

More generally, if λ is a real root of order m, then m corresponding linearly independent solutions are

$$e^{\lambda x}, \quad xe^{\lambda x}, \quad x^2e^{\lambda x}, \quad \cdots, \quad x^{m-1}e^{\lambda x}. \tag{7}$$

We derive these solutions after the next example and indicate how to prove their linear independence.

EXAMPLE 3 Real Double and Triple Roots

Solve the ODE $y^v - 3y^{iv} + 3y''' - y'' = 0$.

Solution. The characteristic equation $\lambda^5 - 3\lambda^4 + 3\lambda^3 - \lambda^2 = 0$ has the roots $\lambda_1 = \lambda_2 = 0$, and $\lambda_3 = \lambda_4 = \lambda_5 = 1$, and the answer is

$$y = c_1 + c_2x + (c_3 + c_4x + c_5x^2)e^x. \tag{8}$$ ■

***Derivation of* (7).** We write the left side of (1) as

$$L[y] = y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_0y.$$

Let $y = e^{\lambda x}$. Then by performing the differentiations we have

$$L[e^{\lambda x}] = (\lambda^n + a_{n-1}\lambda^{n-1} + \cdots + a_0)e^{\lambda x}.$$

Now let λ_1 be a root of mth order of the polynomial on the right, where $m \leqq n$. For $m < n$ let $\lambda_{m+1}, \cdots, \lambda_n$ be the other roots, all different from λ_1. Writing the polynomial in product form, we then have

$$L[e^{\lambda x}] = (\lambda - \lambda_1)^m h(\lambda) e^{\lambda x}$$

with $h(\lambda) = 1$ if $m = n$, and $h(\lambda) = (\lambda - \lambda_{m+1}) \cdots (\lambda - \lambda_n)$ if $m < n$. Now comes the key idea: We differentiate on both sides with respect to λ,

$$\frac{\partial}{\partial \lambda} L[e^{\lambda x}] = m(\lambda - \lambda_1)^{m-1} h(\lambda) e^{\lambda x} + (\lambda - \lambda_1)^m \frac{\partial}{\partial \lambda}[h(\lambda) e^{\lambda x}]. \tag{9}$$

The differentiations with respect to x and λ are independent and the resulting derivatives are continuous, so that we can interchange their order on the left:

$$\frac{\partial}{\partial \lambda} L[e^{\lambda x}] = L\left[\frac{\partial}{\partial \lambda} e^{\lambda x}\right] = L[xe^{\lambda x}]. \tag{10}$$

The right side of (9) is zero for $\lambda = \lambda_1$ because of the factors $\lambda - \lambda_1$ (and $m \geqq 2$ since we have a multiple root!). Hence $L[xe^{\lambda_1 x}] = 0$ by (9) and (10). This proves that $xe^{\lambda_1 x}$ is a solution of (1).

We can repeat this step and produce $x^2 e^{\lambda_1 x}, \cdots, x^{m-1} e^{\lambda_1 x}$ by another $m - 2$ such differentiations with respect to λ. Going one step further would no longer give zero on the right because the lowest power of $\lambda - \lambda_1$ would then be $(\lambda - \lambda_1)^0$, multiplied by $m! h(\lambda)$ and $h(\lambda_1) \neq 0$ because $h(\lambda)$ has no factors $\lambda - \lambda_1$; so we get *precisely* the solutions in (7).

We finally show that the solutions (7) are linearly independent. For a specific n this can be seen by calculating their Wronskian, which turns out to be nonzero. For arbitrary m we can pull out the exponential functions from the Wronskian. This gives $(e^{\lambda x})^m = e^{\lambda m x}$ times a determinant which by "row operations" can be reduced to the Wronskian of 1, $x, \cdots, x^{m-1}$. The latter is constant and different from zero (equal to $1!2! \cdots (m-1)!$). These functions are solutions of the ODE $y^{(m)} = 0$, so that linear independence follows from Theroem 3 in Sec. 3.1.

Multiple Complex Roots

In this case, real solutions are obtained as for complex simple roots above. Consequently, if $\lambda = \gamma + i\omega$ is a **complex double root**, so is the conjugate $\bar{\lambda} = \gamma - i\omega$. Corresponding linearly independent solutions are

$$e^{\gamma x} \cos \omega x, \quad e^{\gamma x} \sin \omega x, \quad xe^{\gamma x} \cos \omega x, \quad xe^{\gamma x} \sin \omega x. \tag{11}$$

The first two of these result from $e^{\lambda x}$ and $e^{\bar{\lambda} x}$ as before, and the second two from $xe^{\lambda x}$ and $xe^{\bar{\lambda} x}$ in the same fashion. Obviously, the corresponding general solution is

$$y = e^{\gamma x}[(A_1 + A_2 x) \cos \omega x + (B_1 + B_2 x) \sin \omega x]. \tag{12}$$

For *complex triple roots* (which hardly ever occur in applications), one would obtain two more solutions $x^2 e^{\gamma x} \cos \omega x$, $x^2 e^{\gamma x} \sin \omega x$, and so on.

PROBLEM SET 3.2

1–6 GENERAL SOLUTION

Solve the given ODE. Show the details of your work.

1. $y''' + 25y' = 0$
2. $y^{iv} + 2y'' + y = 0$
3. $y^{iv} + 4y'' = 0$
4. $(D^3 - D^2 - D + I)y = 0$
5. $(D^4 + 10D^2 + 9I)y = 0$
6. $(D^5 + 8D^3 + 16D)y = 0$

7–13 INITIAL VALUE PROBLEM

Solve the IVP by a CAS, giving a general solution and the particular solution and its graph.

7. $y''' + 3.2y'' + 4.81y' = 0, \quad y(0) = 3.4, \quad y'(0) = -4.6, \quad y''(0) = 9.91$
8. $y''' + 7.5y'' + 14.25y' - 9.125y = 0, \quad y(0) = 10.05, \quad y'(0) = -54.975, \quad y''(0) = 257.5125$
9. $4y''' + 8y'' + 41y' + 37y = 0, \quad y(0) = 9, \quad y'(0) = -6.5, \quad y''(0) = -39.75$
10. $y^{iv} + 4y = 0, \quad y(0) = \frac{1}{2}, \quad y'(0) = -\frac{3}{2}, \quad y''(0) = \frac{5}{2}, \quad y'''(0) = -\frac{7}{2}$
11. $y^{iv} - 9y'' - 400y = 0, \quad y(0) = 0, \quad y'(0) = 0, \quad y''(0) = 41, \quad y'''(0) = 0$
12. $y^{v} - 5y''' + 4y' = 0, \quad y(0) = 3, \quad y'(0) = -5, \quad y''(0) = 11, \quad y'''(0) = -23, \quad y^{iv}(0) = 47$
13. $y^{iv} + 0.45y''' - 0.165y'' + 0.0045y' - 0.00175y = 0, \quad y(0) = 17.4, \quad y'(0) = -2.82, \quad y''(0) = 2.0485, \quad y'''(0) = -1.458675$

14. **PROJECT. Reduction of Order.** This is of practical interest since a single solution of an ODE can often be guessed. For second order, see Example 7 in Sec. 2.1.

    (a) How could you reduce the order of a linear constant-coefficient ODE if a solution is known?

    (b) Extend the method to a variable-coefficient ODE

    $$y''' + p_2(x)y'' + p_1(x)y' + p_0(x)y = 0.$$

    Assuming a solution y_1 to be known, show that another solution is $y_2(x) = u(x)y_1(x)$ with $u(x) = \int z(x)\,dx$ and z obtained by solving

    $$y_1 z'' + (3y_1' + p_2 y_1)z' + (3y_1'' + 2p_2 y_1' + p_1 y_1)z = 0.$$

    (c) Reduce

    $$x^3 y''' - 3x^2 y'' + (6 - x^2)xy' - (6 - x^2)y = 0,$$

    using $y_1 = x$ (perhaps obtainable by inspection).

15. **CAS EXPERIMENT. Reduction of Order.** Starting with a basis, find third-order linear ODEs with variable coefficients for which the reduction to second order turns out to be relatively simple.

3.3 Nonhomogeneous Linear ODEs

We now turn from homogeneous to nonhomogeneous linear ODEs of nth order. We write them in standard form

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x) \qquad \textbf{(1)}$$

with $y^{(n)} = d^n y/dx^n$ as the first term, and $r(x) \not\equiv 0$. As for second-order ODEs, a general solution of (1) on an open interval I of the x-axis is of the form

$$y(x) = y_h(x) + y_p(x). \qquad (2)$$

Here $y_h(x) = c_1 y_1(x) + \cdots + c_n y_n(x)$ is a general solution of the corresponding homogeneous ODE

$$y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0 \qquad (3)$$

on I. Also, y_p is any solution of (1) on I containing no arbitrary constants. If (1) has continuous coefficients and a continuous $r(x)$ on I, then a general solution of (1) exists and includes all solutions. Thus (1) has no singular solutions.

An **initial value problem** for (1) consists of (1) and n **initial conditions**

(4) $$y(x_0) = K_0, \quad y'(x_0) = K_1, \quad \cdots, \quad y^{(n-1)}(x_0) = K_{n-1}$$

with x_0 in I. Under those continuity assumptions it has a unique solution. The ideas of proof are the same as those for $n = 2$ in Sec. 2.7.

Method of Undetermined Coefficients

Equation (2) shows that for solving (1) we have to determine a particular solution of (1). For a constant-coefficient equation

(5) $$y^{(n)} + a_{n-1}y^{(n-1)} + \cdots + a_1 y' + a_0 y = r(x)$$

($a_0, \cdots, a_{n-1}$ constant) and special $r(x)$ as in Sec. 2.7, such a $y_p(x)$ can be determined by the **method of undetermined coefficients,** as in Sec. 2.7, using the following rules.

(A) Basic Rule as in Sec. 2.7.

(B) Modification Rule. *If a term in your choice for* $y_p(x)$ *is a solution of the homogeneous equation* (3), *then multiply this term by* x^k, *where* k *is the smallest positive integer such that this term times* x^k *is not a solution of* (3).

(C) Sum Rule as in Sec. 2.7.

The practical application of the method is the same as that in Sec. 2.7. It suffices to illustrate the typical steps of solving an initial value problem and, in particular, the new Modification Rule, which includes the old Modification Rule as a particular case (with $k = 1$ or 2). We shall see that the technicalities are the same as for $n = 2$, except perhaps for the more involved determination of the constants.

EXAMPLE 1 Initial Value Problem. Modification Rule

Solve the initial value problem

(6) $$y''' + 3y'' + 3y' + y = 30e^{-x}, \qquad y(0) = 3, \qquad y'(0) = -3, \qquad y''(0) = -47.$$

Solution. ***Step 1.*** The characteristic equation is $\lambda^3 + 3\lambda^2 + 3\lambda + 1 = (\lambda + 1)^3 = 0$. It has the triple root $\lambda = -1$. Hence a general solution of the homogeneous ODE is

$$\begin{aligned} y_h &= c_1e^{-x} + c_2xe^{-x} + c_3x^2e^{-x} \\ &= (c_1 + c_2x + c_3x^2)e^{-x}. \end{aligned}$$

Step 2. If we try $y_p = Ce^{-x}$, we get $-C + 3C - 3C + C = 30$, which has no solution. Try Cxe^{-x} and Cx^2e^{-x}. The Modification Rule calls for

$$y_p = Cx^3e^{-x}.$$

Then

$$\begin{aligned} y_p' &= C(3x^2 - x^3)e^{-x}, \\ y_p'' &= C(6x - 6x^2 + x^3)e^{-x}, \\ y_p''' &= C(6 - 18x + 9x^2 - x^3)e^{-x}. \end{aligned}$$

Substitution of these expressions into (6) and omission of the common factor e^{-x} gives

$$C(6 - 18x + 9x^2 - x^3) + 3C(6x - 6x^2 + x^3) + 3C(3x^2 - x^3) + Cx^3 = 30.$$

The linear, quadratic, and cubic terms drop out, and $6C = 30$. Hence $C = 5$. This gives $y_p = 5x^3e^{-x}$.

Step 3. We now write down $y = y_h + y_p$, the general solution of the given ODE. From it we find c_1 by the first initial condition. We insert the value, differentiate, and determine c_2 from the second initial condition, insert the value, and finally determine c_3 from $y''(0)$ and the third initial condition:

$$\begin{aligned} y &= y_h + y_p = (c_1 + c_2x + c_3x^2)e^{-x} + 5x^3e^{-x}, && y(0) = c_1 = 3 \\ y' &= [-3 + c_2 + (-c_2 + 2c_3)x + (15 - c_3)x^2 - 5x^3]e^{-x}, && y'(0) = -3 + c_2 = -3, \quad c_2 = 0 \\ y'' &= [3 + 2c_3 + (30 - 4c_3)x + (-30 + c_3)x^2 + 5x^3]e^{-x}, && y''(0) = 3 + 2c_3 = -47, \quad c_3 = -25. \end{aligned}$$

Hence the *answer* to our problem is (Fig. 73)

$$y = (3 - 25x^2)e^{-x} + 5x^3e^{-x}.$$

The curve of y begins at (0, 3) with a negative slope, as expected from the initial values, and approaches zero as $x \to \infty$. The dashed curve in Fig. 74 is y_p. ■

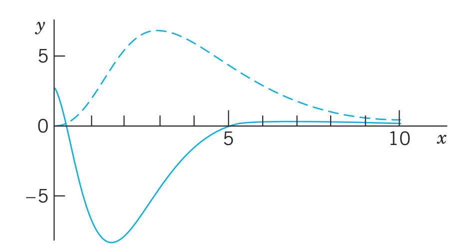


Fig. 74. y and y_p (dashed) in Example 1

Method of Variation of Parameters

The method of variation of parameters (see Sec. 2.10) also extends to arbitrary order n. It gives a particular solution y_p for the nonhomogeneous equation (1) (in standard form with $y^{(n)}$ as the first term!) by the formula

(7)
$$\begin{aligned} y_p(x) &= \sum_{k=1}^{n} y_k(x) \int \frac{W_k(x)}{W(x)} r(x)\, dx \\ &= y_1(x) \int \frac{W_1(x)}{W(x)} r(x)\, dx + \cdots + y_n(x) \int \frac{W_n(x)}{W(x)} r(x)\, dx \end{aligned}$$

on an open interval I on which the coefficients of (1) and $r(x)$ are continuous. In (7) the functions $y_1, \cdots, y_n$ form a basis of the homogeneous ODE (3), with Wronskian W, and W_j ($j = 1, \cdots, n$) is obtained from W by replacing the jth column of W by the column $[0 \quad 0 \quad \cdots \quad 0 \quad 1]^\mathsf{T}$. Thus, when $n = 2$, this becomes identical with (2) in Sec. 2.10,

$$W = \begin{vmatrix} y_1 & y_2 \\ y_1' & y_2' \end{vmatrix}, \qquad W_1 = \begin{vmatrix} 0 & y_2 \\ 1 & y_2' \end{vmatrix} = -y_2, \qquad W_2 = \begin{vmatrix} y_1 & 0 \\ y_1' & 1 \end{vmatrix} = y_1.$$

The proof of (7) uses an extension of the idea of the proof of (2) in Sec. 2.10 and can be found in Ref [A11] listed in App. 1.

EXAMPLE 2 Variation of Parameters. Nonhomogeneous Euler–Cauchy Equation

Solve the nonhomogeneous Euler–Cauchy equation

$$x^3y''' - 3x^2y'' + 6xy' - 6y = x^4 \ln x \qquad (x > 0).$$

Solution. ***Step 1. General solution of the homogeneous ODE.*** Substitution of $y = x^m$ and the derivatives into the homogeneous ODE and deletion of the factor x^m gives

$$m(m-1)(m-2) - 3m(m-1) + 6m - 6 = 0.$$

The roots are 1, 2, 3 and give as a basis

$$y_1 = x, \qquad y_2 = x^2, \qquad y_3 = x^3.$$

Hence the corresponding general solution of the homogeneous ODE is

$$y_h = c_1x + c_2x^2 + c_3x^3.$$

***Step 2. Determinants needed in* (7).** These are

$$W = \begin{vmatrix} x & x^2 & x^3 \\ 1 & 2x & 3x^2 \\ 0 & 2 & 6x \end{vmatrix} = 2x^3$$

$$W_1 = \begin{vmatrix} 0 & x^2 & x^3 \\ 0 & 2x & 3x^2 \\ 1 & 2 & 6x \end{vmatrix} = x^4$$

$$W_2 = \begin{vmatrix} x & 0 & x^3 \\ 1 & 0 & 3x^2 \\ 0 & 1 & 6x \end{vmatrix} = -2x^3$$

$$W_3 = \begin{vmatrix} x & x^2 & 0 \\ 1 & 2x & 0 \\ 0 & 2 & 1 \end{vmatrix} = x^2.$$

Step 3. Integration. In (7) we also need the right side $r(x)$ of our ODE in standard form, obtained by division of the given equation by the coefficient x^3 of y'''; thus, $r(x) = (x^4 \ln x)/x^3 = x \ln x$. In (7) we have the simple quotients $W_1/W = x/2$, $W_2/W = -1$, $W_3/W = 1/(2x)$. Hence (7) becomes

$$\begin{aligned} y_p &= x \int \frac{x}{2} x \ln x \, dx - x^2 \int x \ln x \, dx + x^3 \int \frac{1}{2x} x \ln x \, dx \\ &= \frac{x}{2}\left(\frac{x^3}{3} \ln x - \frac{x^3}{9}\right) - x^2\left(\frac{x^2}{2} \ln x - \frac{x^2}{4}\right) + \frac{x^3}{2}(x \ln x - x). \end{aligned}$$

Simplification gives $y_p = \frac{1}{6}x^4 (\ln x - \frac{11}{6})$. Hence the answer is

$$y = y_h + y_p = c_1x + c_2x^2 + c_3x^3 + \tfrac{1}{6}x^4 (\ln x - \tfrac{11}{6}).$$

Figure 75 shows y_p. Can you explain the shape of this curve? Its behavior near $x = 0$? The occurrence of a minimum? Its rapid increase? Why would the method of undetermined coefficients not have given the solution? ■

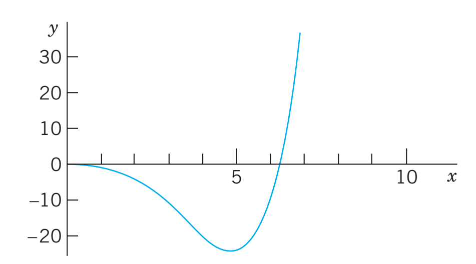


Fig. 75. Particular solution y_p of the nonhomogeneous Euler–Cauchy equation in Example 2

Application: Elastic Beams

Whereas second-order ODEs have various applications, of which we have discussed some of the more important ones, higher order ODEs have much fewer engineering applications. An important fourth-order ODE governs the bending of elastic beams, such as wooden or iron girders in a building or a bridge.

A related application of vibration of beams does not fit in here since it leads to PDEs and will therefore be discussed in Sec. 12.3.

EXAMPLE 3 Bending of an Elastic Beam under a Load

We consider a beam B of length L and constant (e.g., **rectangular**) cross section and homogeneous elastic material (e.g., steel); see Fig. 76. We assume that under its own weight the beam is bent so little that it is practically straight. If we apply a load to B in a vertical plane through the axis of symmetry (the x-axis in Fig. 76), B is bent. Its axis is curved into the so-called **elastic curve** C (or **deflection curve**). It is shown in elasticity theory that the bending moment $M(x)$ is proportional to the curvature $k(x)$ of C. We assume the bending to be small, so that the deflection $y(x)$ and its derivative $y'(x)$ (determining the tangent direction of C) are small. Then, by calculus, $k = y''/(1 + y'^2)^{3/2} \approx y''$. Hence

$$M(x) = EIy''(x).$$

EI is the constant of proportionality. E is *Young's modulus of elasticity* of the material of the beam. I is the moment of inertia of the cross section about the (horizontal) z-axis in Fig. 76.

Elasticity theory shows further that $M''(x) = f(x)$, where $f(x)$ is the load per unit length. Together,

$$EIy^{\text{iv}} = f(x). \tag{8}$$

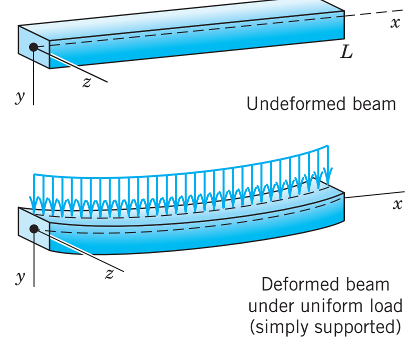


Fig. 76. Elastic beam

In applications the most important supports and corresponding boundary conditions are as follows and shown in Fig. 77.

(A) Simply supported $\qquad y = y'' = 0$ at $x = 0$ and L

(B) Clamped at both ends $\qquad y = y' = 0$ at $x = 0$ and L

(C) Clamped at $x = 0$, free at $x = L$ $\qquad y(0) = y'(0) = 0, y''(L) = y'''(L) = 0.$

The boundary condition $y = 0$ means no displacement at that point, $y' = 0$ means a horizontal tangent, $y'' = 0$ means no bending moment, and $y''' = 0$ means no shear force.

Let us apply this to the uniformly loaded simply supported beam in Fig. 76. The load is $f(x) \equiv f_0 = \text{const}$. Then (8) is

$$y^{\text{iv}} = k, \qquad k = \frac{f_0}{EI}. \tag{9}$$

This can be solved simply by calculus. Two integrations give

$$y'' = \frac{k}{2}x^2 + c_1 x + c_2.$$

$y''(0) = 0$ gives $c_2 = 0$. Then $y''(L) = L(\frac{1}{2}kL + c_1) = 0$, $c_1 = -kL/2$ (since $L \neq 0$). Hence

$$y'' = \frac{k}{2}(x^2 - Lx).$$

Integrating this twice, we obtain

$$y = \frac{k}{2}\left(\frac{1}{12}x^4 - \frac{L}{6}x^3 + c_3 x + c_4\right)$$

with $c_4 = 0$ from $y(0) = 0$. Then

$$y(L) = \frac{kL}{2}\left(\frac{L^3}{12} - \frac{L^3}{6} + c_3\right) = 0, \qquad c_3 = \frac{L^3}{12}.$$

Inserting the expression for k, we obtain as our solution

$$y = \frac{f_0}{24EI}(x^4 - 2Lx^3 + L^3 x).$$

Since the boundary conditions at both ends are the same, we expect the deflection $y(x)$ to be "symmetric" with respect to $L/2$, that is, $y(x) = y(L - x)$. Verify this directly or set $x = u + L/2$ and show that y becomes an even function of u,

$$y = \frac{f_0}{24EI}\left(u^2 - \frac{1}{4}L^2\right)\left(u^2 - \frac{5}{4}L^2\right).$$

From this we can see that the maximum deflection in the middle at $u = 0\,(x = L/2)$ is $5f_0L^4/(16 \cdot 24EI)$. Recall that the positive direction points downward. ■

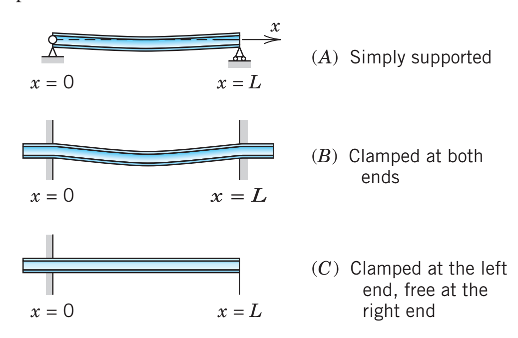


Fig. 77. Supports of a beam

PROBLEM SET 3.3

1–7 GENERAL SOLUTION

Solve the following ODEs, showing the details of your work.

1. $y''' + 3y'' + 3y' + y = e^x - x - 1$
2. $y''' + 2y'' - y' - 2y = 1 - 4x^3$
3. $(D^4 + 10D^2 + 9I)y = 6.5 \sinh 2x$
4. $(D^3 + 3D^2 - 5D - 39I)y = -300 \cos x$
5. $(x^3D^3 + x^2D^2 - 2xD + 2I)y = x^{-2}$
6. $(D^3 + 4D)y = \sin x$
7. $(D^3 - 9D^2 + 27D - 27I)y = 27 \sin 3x$

8–13 INITIAL VALUE PROBLEM

Solve the given IVP, showing the details of your work.

8. $y^{\text{iv}} - 5y'' + 4y = 10e^{-3x}$, $y(0) = 1$, $y'(0) = 0$, $y''(0) = 0$, $y'''(0) = 0$
9. $y^{\text{iv}} + 5y'' + 4y = 90 \sin 4x$, $y(0) = 1$, $y'(0) = 2$, $y''(0) = -1$, $y'''(0) = -32$
10. $x^3y''' + xy' - y = x^2$, $y(1) = 1$, $y'(1) = 3$, $y''(1) = 14$
11. $(D^3 - 2D^2 - 3D)y = 74e^{-3x} \sin x$, $y(0) = -1.4$, $y'(0) = 3.2$, $y''(0) = -5.2$
12. $(D^3 - 2D^2 - 9D + 18I)y = e^{2x}$, $y(0) = 4.5$, $y'(0) = 8.8$, $y''(0) = 17.2$
13. $(D^3 - 4D)y = 10 \cos x + 5 \sin x$, $y(0) = 3$, $y'(0) = -2$, $y''(0) = -1$

14. **CAS EXPERIMENT. Undetermined Coefficients.** Since variation of parameters is generally complicated, it seems worthwhile to try to extend the other method. Find out experimentally for what ODEs this is possible and for what not. *Hint:* Work backward, solving ODEs with a CAS and then looking whether the solution could be obtained by undetermined coefficients. For example, consider

$$y''' - 3y'' + 3y' - y = x^{1/2}e^x$$

and

$$x^3y''' + x^2y'' - 2xy' + 2y = x^3 \ln x.$$

15. **WRITING REPORT. Comparison of Methods.** Write a report on the method of undetermined coefficients and the method of variation of parameters, discussing and comparing the advantages and disadvantages of each method. Illustrate your findings with typical examples. Try to show that the method of undetermined coefficients, say, for a third-order ODE with constant coefficients and an exponential function on the right, can be derived from the method of variation of parameters.

CHAPTER 3 REVIEW QUESTIONS AND PROBLEMS

1. What is the superposition or linearity principle? For what nth-order ODEs does it hold?
2. List some other basic theorems that extend from second-order to nth-order ODEs.
3. If you know a general solution of a homogeneous linear ODE, what do you need to obtain from it a general solution of a corresponding nonhomogeneous linear ODE?
4. What form does an initial value problem for an nth-order linear ODE have?
5. What is the Wronskian? What is it used for?

6–15 GENERAL SOLUTION

Solve the given ODE. Show the details of your work.

6. $y^{\text{iv}} - 3y'' - 4y = 0$
7. $y''' + 4y'' + 13y' = 0$
8. $y''' - 4y'' - y' + 4y = 30e^{2x}$
9. $(D^4 - 16I)y = -15 \cosh x$
10. $x^2y''' + 3xy'' - 2y' = 0$
11. $y''' + 4.5y'' + 6.75y' + 3.375y = 0$
12. $(D^3 - D)y = \sinh 0.8x$
13. $(D^3 + 6D^2 + 12D + 8I)y = 8x^2$
14. $(D^4 - 13D^2 + 36I)y = 12e^x$
15. $4x^3y''' + 3xy' - 3y = 10$

16–20 INITIAL VALUE PROBLEM

Solve the IVP. Show the details of your work.

16. $(D^3 - D^2 - D + I)y = 0$, $y(0) = 0$, $Dy(0) = 1$, $D^2y(0) = 0$
17. $y''' + 5y'' + 24y' + 20y = x$, $y(0) = 1.94$, $y'(0) = -3.95$, $y'' = -24$
18. $(D^4 - 26D^2 + 25I)y = 50(x + 1)^2$, $y(0) = 12.16$, $Dy(0) = -6$, $D^2y(0) = 34$, $D^3y(0) = -130$
19. $(D^3 + 9D^2 + 23D + 15I)y = 12\exp(-4x)$, $y(0) = 9$, $Dy(0) = -41$, $D^2y(0) = 189$
20. $(D^3 + 3D^2 + 3D + I)y = 8 \sin x$, $y(0) = -1$, $y'(0) = -3$, $y''(0) = 5$

SUMMARY OF CHAPTER 3
Higher Order Linear ODEs

Compare with the similar Summary of Chap. 2 (the case n = 2).
Chapter 3 extends Chap. 2 from order $n = 2$ to arbitrary order n. An ***n*th-order linear ODE** is an ODE that can be written

$$(1) \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = r(x)$$

with $y^{(n)} = d^n y/dx^n$ as the first term; we again call this the **standard form**. Equation (1) is called **homogeneous** if $r(x) \equiv 0$ on a given open interval I considered, **nonhomogeneous** if $r(x) \not\equiv 0$ on I. For the homogeneous ODE

$$(2) \qquad y^{(n)} + p_{n-1}(x)y^{(n-1)} + \cdots + p_1(x)y' + p_0(x)y = 0$$

the ***superposition principle*** (Sec. 3.1) holds, just as in the case $n = 2$. A **basis** or **fundamental system** of solutions of (2) on I consists of n linearly independent solutions $y_1, \cdots, y_n$ of (2) on I. A **general solution** of (2) on I is a linear combination of these,

$$(3) \qquad y = c_1y_1 + \cdots + c_ny_n \qquad (c_1, \cdots, c_n \text{ arbitrary constants}).$$

A **general solution** of the nonhomogeneous ODE (1) on I is of the form

$$(4) \qquad y = y_h + y_p \qquad \text{(Sec. 3.3)}.$$

Here, y_p is a particular solution of (1) and is obtained by two methods (***undetermined coefficients*** or ***variation of parameters***) explained in Sec. 3.3.

An **initial value problem** for (1) or (2) consists of one of these ODEs and n initial conditions (Secs. 3.1, 3.3)

$$(5) \qquad y(x_0) = K_0, \qquad y'(x_0) = K_1, \qquad \cdots, \qquad y^{(n-1)}(x_0) = K_{n-1}$$

with given x_0 in I and given $K_0, \cdots, K_{n-1}$. If $p_0, \cdots, p_{n-1}, r$ are continuous on I, then general solutions of (1) and (2) on I exist, and initial value problems (1), (5) or (2), (5) have a unique solution.

Systems of ODEs. Phase Plane. Qualitative Methods

Tying in with Chap. 3, we present another method of solving higher order ODEs in Sec. 4.1. This converts any *n*th-order ODE into a system of *n* first-order ODEs. We also show some applications. Moreover, in the same section we solve systems of first-order ODEs that occur directly in applications, that is, not derived from an *n*th-order ODE but dictated by the application such as two tanks in mixing problems and two circuits in electrical networks. (The elementary aspects of vectors and matrices needed in this chapter are reviewed in Sec. 4.0 and are probably familiar to most students.)

In Sec. 4.3 we introduce a totally different way of looking at systems of ODEs. The method consists of examining the general behavior of whole families of solutions of ODEs in the ***phase plane***, and aptly is called the phase plane method. It gives information on the **stability** of solutions. (*Stability of a physical system* is desirable and means roughly that a small change at some instant causes only a small change in the behavior of the system at later times.) This approach to systems of ODEs is a **qualitative method** because it depends only on the nature of the ODEs and does not require the actual solutions. This can be very useful because it is often difficult or even impossible to solve systems of ODEs. In contrast, the approach of actually solving a system is known as a *quantitative method*.

The phase plane method has many applications in control theory, circuit theory, population dynamics and so on. Its use in linear systems is discussed in Secs. 4.3, 4.4, and 4.6 and its even more important use in nonlinear systems is discussed in Sec. 4.5 with applications to the pendulum equation and the Lokta–Volterra population model. The chapter closes with a discussion of nonhomogeneous linear systems of ODEs.

NOTATION. We continue to denote unknown functions by y; thus, $y_1(t), y_2(t)$— analogous to Chaps. 1–3. (Note that some authors use x for functions, $x_1(t), x_2(t)$ when dealing with systems of ODEs.)

Prerequisite: Chap. 2.
References and Answers to Problems: App. 1 Part A, and App. 2.

4.0 For Reference: Basics of Matrices and Vectors

For clarity and simplicity of notation, we use matrices and vectors in our discussion of linear systems of ODEs. We need only a few elementary facts (and not the bulk of the material of Chaps. 7 and 8). Most students will very likely be already familiar

with these facts. Thus ***this section is for reference only***. *Begin with Sec. 4.1 and consult 4.0 as needed.*

Most of our linear systems will consist of two linear ODEs in two unknown functions $y_1(t)$, $y_2(t)$,

$$(1)\qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2, \\ y_2' &= a_{21}y_1 + a_{22}y_2, \end{aligned} \qquad \text{for example,} \qquad \begin{aligned} y_1' &= -5y_1 + 2y_2 \\ y_2' &= \ \ 13y_1 + \tfrac{1}{2}y_2 \end{aligned}$$

(perhaps with additional *given* functions $g_1(t)$, $g_2(t)$ on the right in the two ODEs).

Similarly, a linear system of n first-order ODEs in n unknown functions $y_1(t), \cdots, y_n(t)$ is of the form

$$(2)\qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + \cdots + a_{1n}y_n \\ y_2' &= a_{21}y_1 + a_{22}y_2 + \cdots + a_{2n}y_n \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ y_n' &= a_{n1}y_1 + a_{n2}y_2 + \cdots + a_{nn}y_n \end{aligned}$$

(perhaps with an additional given function on the right in each ODE).

Some Definitions and Terms

Matrices. In (1) the (constant or variable) coefficients form a **2 × 2 matrix A**, that is, an array

$$(3)\qquad \mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \text{for example,} \qquad \mathbf{A} = \begin{bmatrix} -5 & 2 \\ 13 & \tfrac{1}{2} \end{bmatrix}.$$

Similarly, the coefficients in (2) form an ***n* × *n* matrix**

$$(4)\qquad \mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}.$$

The $a_{11}, a_{12}, \cdots$ are called **entries**, the horizontal lines **rows**, and the vertical lines **columns**. Thus, in (3) the first row is $[a_{11} \quad a_{12}]$, the second row is $[a_{21} \quad a_{22}]$, and the first and second columns are

$$\begin{bmatrix} a_{11} \\ a_{21} \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} a_{12} \\ a_{22} \end{bmatrix}.$$

In the "***double subscript notation***" for entries, the first subscript denotes the *row* and the second the *column* in which the entry stands. Similarly in (4). The **main diagonal** is the diagonal $a_{11} \quad a_{22} \quad \cdots \quad a_{nn}$ in (4), hence $a_{11} \quad a_{22}$ in (3).

We shall need only **square matrices**, that is, matrices with the same number of rows and columns, as in (3) and (4).

Vectors. A **column vector x** with n **components** $x_1, \cdots, x_n$ is of the form

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \qquad \text{thus if } n = 2, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

Similarly, a **row vector v** is of the form

$$\mathbf{v} = [v_1 \quad \cdots \quad v_n], \qquad \text{thus if } n = 2, \text{ then} \qquad \mathbf{v} = [v_1 \quad v_2].$$

Calculations with Matrices and Vectors

Equality. Two $n \times n$ matrices are *equal* if and only if corresponding entries are equal. Thus for $n = 2$, let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{B} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix}.$$

Then $\mathbf{A} = \mathbf{B}$ if and only if

$$\begin{aligned} a_{11} &= b_{11}, & a_{12} &= b_{12} \\ a_{21} &= b_{21}, & a_{22} &= b_{22}. \end{aligned}$$

Two column vectors (or two row vectors) are *equal* if and only if they both have n components and corresponding components are equal. Thus, let

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \qquad \text{Then} \qquad \mathbf{v} = \mathbf{x} \quad \text{if and only if} \quad \begin{aligned} v_1 &= x_1 \\ v_2 &= x_2. \end{aligned}$$

Addition is performed by adding corresponding entries (or components); here, matrices must both be $n \times n$, and vectors must both have the same number of components. Thus for $n = 2$,

$$(5) \qquad \mathbf{A} + \mathbf{B} = \begin{bmatrix} a_{11} + b_{11} & a_{12} + b_{12} \\ a_{21} + b_{21} & a_{22} + b_{22} \end{bmatrix}, \qquad \mathbf{v} + \mathbf{x} = \begin{bmatrix} v_1 + x_1 \\ v_2 + x_2 \end{bmatrix}.$$

Scalar multiplication (multiplication by a number c) is performed by multiplying each entry (or component) by c. For example, if

$$\mathbf{A} = \begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix}, \qquad \text{then} \qquad -7\mathbf{A} = \begin{bmatrix} -63 & -21 \\ 14 & 0 \end{bmatrix}.$$

If

$$\mathbf{v} = \begin{bmatrix} 0.4 \\ -13 \end{bmatrix}, \qquad \text{then} \qquad 10\mathbf{v} = \begin{bmatrix} 4 \\ -130 \end{bmatrix}.$$

Matrix Multiplication. The product $\mathbf{C} = \mathbf{AB}$ (in this order) of two $n \times n$ matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ is the $n \times n$ matrix $\mathbf{C} = [c_{jk}]$ with entries

(6)
$$c_{jk} = \sum_{m=1}^{n} a_{jm} b_{mk} \qquad \begin{matrix} j = 1, \cdots, n \\ k = 1, \cdots, n, \end{matrix}$$

that is, multiply each entry in the jth *row* of **A** by the corresponding entry in the kth *column* of **B** and then add these n products. One says briefly that this is a "multiplication of rows into columns." For example,

$$\begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} 1 & -4 \\ 2 & 5 \end{bmatrix} = \begin{bmatrix} 9 \cdot 1 + 3 \cdot 2 & 9 \cdot (-4) + 3 \cdot 5 \\ -2 \cdot 1 + 0 \cdot 2 & (-2) \cdot (-4) + 0 \cdot 5 \end{bmatrix},$$

$$= \begin{bmatrix} 15 & -21 \\ -2 & 8 \end{bmatrix}.$$

CAUTION! Matrix multiplication is ***not commutative,*** $\mathbf{AB} \neq \mathbf{BA}$ in general. In our example,

$$\begin{bmatrix} 1 & -4 \\ 2 & 5 \end{bmatrix} \begin{bmatrix} 9 & 3 \\ -2 & 0 \end{bmatrix} = \begin{bmatrix} 1 \cdot 9 + (-4) \cdot (-2) & 1 \cdot 3 + (-4) \cdot 0 \\ 2 \cdot 9 + 5 \cdot (-2) & 2 \cdot 3 + 5 \cdot 0 \end{bmatrix}$$

$$= \begin{bmatrix} 17 & 3 \\ 8 & 6 \end{bmatrix}.$$

Multiplication of an $n \times n$ matrix **A** by a vector **x** with n components is defined by the same rule: $\mathbf{v} = \mathbf{Ax}$ is the vector with the n components

$$v_j = \sum_{m=1}^{n} a_{jm} x_m \qquad j = 1, \cdots, n.$$

For example,

$$\begin{bmatrix} 12 & 7 \\ -8 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 12x_1 + 7x_2 \\ -8x_1 + 3x_2 \end{bmatrix}.$$

Systems of ODEs as Vector Equations

Differentiation. The *derivative* of a matrix (or vector) with variable entries (or components) is obtained by differentiating each entry (or component). Thus, if

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix} = \begin{bmatrix} e^{-2t} \\ \sin t \end{bmatrix}, \qquad \text{then} \qquad \mathbf{y}'(t) = \begin{bmatrix} y_1'(t) \\ y_2'(t) \end{bmatrix} = \begin{bmatrix} -2e^{-2t} \\ \cos t \end{bmatrix}.$$

Using matrix multiplication and differentiation, we can now write (1) as

$$\mathbf{y}' = \begin{bmatrix} y_1' \\ y_2' \end{bmatrix} = \mathbf{A}\mathbf{y} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \text{e.g.,} \quad \mathbf{y}' = \begin{bmatrix} -5 & 2 \\ 13 & \frac{1}{2} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{7}$$

Similarly for (2) by means of an $n \times n$ matrix $\mathbf{A}$ and a column vector $\mathbf{y}$ with n components, namely, $\mathbf{y}' = \mathbf{A}\mathbf{y}$. The vector equation (7) is equivalent to two equations for the components, and these are precisely the two ODEs in (1).

Some Further Operations and Terms

Transposition is the operation of writing columns as rows and conversely and is indicated by T. Thus the transpose $\mathbf{A}^{\mathsf{T}}$ of the 2×2 matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} = \begin{bmatrix} -5 & 2 \\ 13 & \frac{1}{2} \end{bmatrix} \quad \text{is} \quad \mathbf{A}^{\mathsf{T}} = \begin{bmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{bmatrix} = \begin{bmatrix} -5 & 13 \\ 2 & \frac{1}{2} \end{bmatrix}.$$

The transpose of a column vector, say,

$$\mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}, \quad \text{is a row vector,} \quad \mathbf{v}^{\mathsf{T}} = [v_1 \quad v_2],$$

and conversely.

Inverse of a Matrix. The $n \times n$ **unit matrix I** is the $n \times n$ matrix with main diagonal $1, 1, \cdots, 1$ and all other entries zero. If, for a given $n \times n$ matrix $\mathbf{A}$, there is an $n \times n$ matrix $\mathbf{B}$ such that $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, then $\mathbf{A}$ is called **nonsingular** and $\mathbf{B}$ is called the **inverse** of $\mathbf{A}$ and is denoted by $\mathbf{A}^{-1}$; thus

$$\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}. \tag{8}$$

The inverse exists if the determinant $\det \mathbf{A}$ of $\mathbf{A}$ is not zero.

If $\mathbf{A}$ has no inverse, it is called **singular**. For $n = 2$,

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}, \tag{9}$$

where the **determinant** of $\mathbf{A}$ is

$$\det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{10}$$

(For general n, see Sec. 7.7, but this will not be needed in this chapter.)

Linear Independence. r given vectors $\mathbf{v}^{(1)}, \cdots, \mathbf{v}^{(r)}$ with n components are called a *linearly independent set* or, more briefly, **linearly independent**, if

$$c_1\mathbf{v}^{(1)} + \cdots + c_r\mathbf{v}^{(r)} = \mathbf{0} \tag{11}$$

implies that all scalars $c_1, \cdots, c_r$ must be zero; here, $\mathbf{0}$ denotes the **zero vector**, whose n components are all zero. If (11) also holds for scalars not all zero (so that at least one of these scalars is not zero), then these vectors are called a *linearly dependent set* or, briefly, **linearly dependent**, because then at least one of them can be expressed as a **linear combination** of the others; that is, if, for instance, $c_1 \neq 0$ in (11), then we can obtain

$$\mathbf{v}^{(1)} = -\frac{1}{c_1}(c_2\mathbf{v}^{(2)} + \cdots + c_r\mathbf{v}^{(r)}).$$

Eigenvalues, Eigenvectors

Eigenvalues and eigenvectors will be very important in this chapter (and, as a matter of fact, throughout mathematics).

Let $\mathbf{A} = [a_{jk}]$ be an $n \times n$ matrix. Consider the equation

(12)
$$\mathbf{Ax} = \lambda\mathbf{x}$$

where λ is a scalar (a real or complex number) to be determined and $\mathbf{x}$ is a vector to be determined. Now, for every λ, a solution is $\mathbf{x} = \mathbf{0}$. A scalar λ such that (12) holds for some vector $\mathbf{x} \neq \mathbf{0}$ is called an **eigenvalue** of $\mathbf{A}$, and this vector is called an **eigenvector** of $\mathbf{A}$ corresponding to this eigenvalue λ.

We can write (12) as $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{0}$ or

(13)
$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

These are n linear algebraic equations in the n unknowns $x_1, \cdots, x_n$ (the components of $\mathbf{x}$). For these equations to have a solution $\mathbf{x} \neq \mathbf{0}$, the determinant of the coefficient matrix $\mathbf{A} - \lambda\mathbf{I}$ must be zero. This is proved as a basic fact in linear algebra (Theorem 4 in Sec. 7.7). In this chapter we need this only for $n = 2$. Then (13) is

(14)
$$\begin{bmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix};$$

in components,

(14*)
$$\begin{aligned} (a_{11} - \lambda)x_1 + \quad a_{12}x_2 \quad &= 0 \\ a_{21}x_1 \quad + (a_{22} - \lambda)x_2 &= 0. \end{aligned}$$

Now $\mathbf{A} - \lambda\mathbf{I}$ is singular if and only if its determinant $\det(\mathbf{A} - \lambda\mathbf{I})$, called the **characteristic determinant** of $\mathbf{A}$ (also for general n), is zero. This gives

(15)
$$\begin{aligned} \det(\mathbf{A} - \lambda\mathbf{I}) &= \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} \\ &= (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} \\ &= \lambda^2 - (a_{11} + a_{22})\lambda + a_{11}a_{22} - a_{12}a_{21} = 0. \end{aligned}$$

This quadratic equation in λ is called the **characteristic equation** of **A**. Its solutions are the eigenvalues λ_1 and λ_2 of **A**. First determine these. Then use (14*) with $\lambda = \lambda_1$ to determine an eigenvector $\mathbf{x}^{(1)}$ of **A** corresponding to λ_1. Finally use (14*) with $\lambda = \lambda_2$ to find an eigenvector $\mathbf{x}^{(2)}$ of **A** corresponding to λ_2. Note that if **x** is an eigenvector of **A**, so is $k\mathbf{x}$ with any $k \neq 0$.

EXAMPLE 1 Eigenvalue Problem

Find the eigenvalues and eigenvectors of the matrix

$$\mathbf{A} = \begin{bmatrix} -4.0 & 4.0 \\ -1.6 & 1.2 \end{bmatrix} \tag{16}$$

Solution. The characteristic equation is the quadratic equation

$$\det |\mathbf{A} - \lambda\mathbf{I}| = \begin{vmatrix} -4 - \lambda & 4 \\ -1.6 & 1.2 - \lambda \end{vmatrix} = \lambda^2 + 2.8\lambda + 1.6 = 0.$$

It has the solutions $\lambda_1 = -2$ and $\lambda_2 = -0.8$. These are the eigenvalues of **A**.

Eigenvectors are obtained from (14*). For $\lambda = \lambda_1 = -2$ we have from (14*)

$$\begin{aligned} (-4.0 + 2.0)x_1 + \quad 4.0x_2 \quad &= 0 \\ -1.6x_1 \quad + (1.2 + 2.0)x_2 &= 0. \end{aligned}$$

A solution of the first equation is $x_1 = 2, x_2 = 1$. This also satisfies the second equation. (Why?) Hence an eigenvector of **A** corresponding to $\lambda_1 = -2.0$ is

$$\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}. \qquad \text{Similarly,} \qquad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix} \tag{17}$$

is an eigenvector of **A** corresponding to $\lambda_2 = -0.8$, as obtained from (14*) with $\lambda = \lambda_2$. Verify this. ■

4.1 Systems of ODEs as Models in Engineering Applications

We show how systems of ODEs are of practical importance as follows. We first illustrate how systems of ODEs can serve as models in various applications. Then we show how a higher order ODE (with the highest derivative standing alone on one side) can be reduced to a first-order system.

EXAMPLE 1 Mixing Problem Involving Two Tanks

A mixing problem involving a single tank is modeled by a single ODE, and you may first review the corresponding Example 3 in Sec. 1.3 because the principle of modeling will be the same for two tanks. The model will be a system of two first-order ODEs.

Tank T_1 and T_2 in Fig. 78 contain initially 100 gal of water each. In T_1 the water is pure, whereas 150 lb of fertilizer are dissolved in T_2. By circulating liquid at a rate of 2 gal/min and stirring (to keep the mixture uniform) the amounts of fertilizer $y_1(t)$ in T_1 and $y_2(t)$ in T_2 change with time t. How long should we let the liquid circulate so that T_1 will contain at least half as much fertilizer as there will be left in T_2?

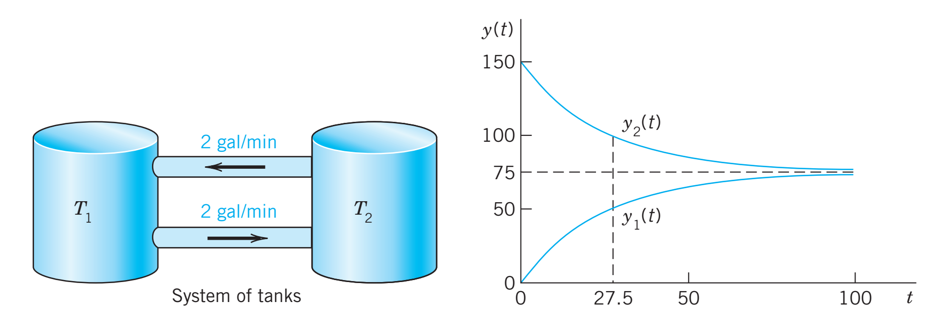


Fig. 78. Fertilizer content in Tanks T_1 (lower curve) and T_2

Solution. ***Step 1.*** **Setting up the model.** As for a single tank, the time rate of change $y_1'(t)$ of $y_1(t)$ equals inflow minus outflow. Similarly for tank T_2. From Fig. 78 we see that

$$y_1' = \text{Inflow/min} - \text{Outflow/min} = \frac{2}{100}y_2 - \frac{2}{100}y_1 \qquad \text{(Tank } T_1\text{)}$$

$$y_2' = \text{Inflow/min} - \text{Outflow/min} = \frac{2}{100}y_1 - \frac{2}{100}y_2 \qquad \text{(Tank } T_2\text{).}$$

Hence the mathematical model of our mixture problem is the system of first-order ODEs

$$\begin{aligned} y_1' &= -0.02y_1 + 0.02y_2 \qquad &\text{(Tank } T_1\text{)} \\ y_2' &= 0.02y_1 - 0.02y_2 \qquad &\text{(Tank } T_2\text{).} \end{aligned}$$

As a vector equation with column vector $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$ and matrix $\mathbf{A}$ this becomes

$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \qquad \text{where} \qquad \mathbf{A} = \begin{bmatrix} -0.02 & 0.02 \\ 0.02 & -0.02 \end{bmatrix}.$$

Step 2. **General solution.** As for a single equation, we try an exponential function of t,

$$\text{(1)} \qquad \mathbf{y} = \mathbf{x}e^{\lambda t}. \qquad \text{Then} \qquad \mathbf{y}' = \lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}.$$

Dividing the last equation $\lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}$ by $e^{\lambda t}$ and interchanging the left and right sides, we obtain

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

We need nontrivial solutions (solutions that are not identically zero). Hence we have to look for eigenvalues and eigenvectors of $\mathbf{A}$. The eigenvalues are the solutions of the characteristic equation

$$\text{(2)} \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -0.02 - \lambda & 0.02 \\ 0.02 & -0.02 - \lambda \end{vmatrix} = (-0.02 - \lambda)^2 - 0.02^2 = \lambda(\lambda + 0.04) = 0.$$

We see that $\lambda_1 = 0$ (which can very well happen—don't get mixed up—it is eigen*vectors* that must not be zero) and $\lambda_2 = -0.04$. Eigenvectors are obtained from (14*) in Sec. 4.0 with $\lambda = 0$ and $\lambda = -0.04$. For our present $\mathbf{A}$ this gives [we need only the first equation in (14*)]

$$-0.02x_1 + 0.02x_2 = 0 \qquad \text{and} \qquad (-0.02 + 0.04)x_1 + 0.02x_2 = 0,$$

respectively. Hence $x_1 = x_2$ and $x_1 = -x_2$, respectively, and we can take $x_1 = x_2 = 1$ and $x_1 = -x_2 = 1$. This gives two eigenvectors corresponding to $\lambda_1 = 0$ and $\lambda_2 = -0.04$, respectively, namely,

$$\mathbf{x}^{(1)} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \quad \text{and} \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

From (1) and the superposition principle (which continues to hold for systems of homogeneous linear ODEs) we thus obtain a solution

$$\mathbf{y} = c_1\mathbf{x}^{(1)}e^{\lambda_1 t} + c_2\mathbf{x}^{(2)}e^{\lambda_2 t} = c_1\begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2\begin{bmatrix} 1 \\ -1 \end{bmatrix}e^{-0.04t} \tag{3}$$

where c_1 and c_2 are arbitrary constants. Later we shall call this a **general solution**.

***Step 3.* Use of initial conditions.** The initial conditions are $y_1(0) = 0$ (no fertilizer in tank T_1) and $y_2(0) = 150$. From this and (3) with $t = 0$ we obtain

$$\mathbf{y}(0) = c_1\begin{bmatrix} 1 \\ 1 \end{bmatrix} + c_2\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 \\ c_1 - c_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 150 \end{bmatrix}.$$

In components this is $c_1 + c_2 = 0$, $c_1 - c_2 = 150$. The solution is $c_1 = 75$, $c_2 = -75$. This gives the answer

$$\mathbf{y} = 75\mathbf{x}^{(1)} - 75\mathbf{x}^{(2)}e^{-0.04t} = 75\begin{bmatrix} 1 \\ 1 \end{bmatrix} - 75\begin{bmatrix} 1 \\ -1 \end{bmatrix}e^{-0.04t}.$$

In components,

$$y_1 = 75 - 75e^{-0.04t} \qquad \text{(Tank } T_1\text{, lower curve)}$$

$$y_2 = 75 + 75e^{-0.04t} \qquad \text{(Tank } T_2\text{, upper curve).}$$

Figure 78 shows the exponential increase of y_1 and the exponential decrease of y_2 to the common limit 75 lb. Did you expect this for physical reasons? Can you physically explain why the curves look "symmetric"? Would the limit change if T_1 initially contained 100 lb of fertilizer and T_2 contained 50 lb?

***Step 4.* Answer.** T_1 contains half the fertilizer amount of T_2 if it contains 1/3 of the total amount, that is, 50 lb. Thus

$$y_1 = 75 - 75e^{-0.04t} = 50, \qquad e^{-0.04t} = \tfrac{1}{3}, \qquad t = (\ln 3)/0.04 = 27.5.$$

Hence the fluid should circulate for at least about half an hour. ■

EXAMPLE 2 Electrical Network

Find the currents $I_1(t)$ and $I_2(t)$ in the network in Fig. 79. Assume all currents and charges to be zero at $t = 0$, the instant when the switch is closed.

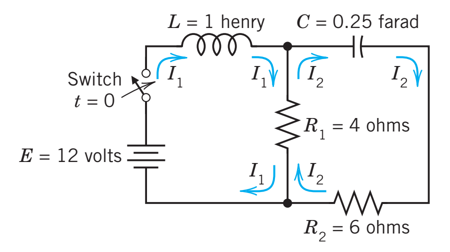


Fig. 79. Electrical network in Example 2

Solution. ***Step 1.* Setting up the mathematical model.** The model of this network is obtained from Kirchhoff's Voltage Law, as in Sec. 2.9 (where we considered single circuits). Let $I_1(t)$ and $I_2(t)$ be the currents

in the left and right loops, respectively. In the left loop, the voltage drops are $LI_1' = I_1'$ [V] over the inductor and $R_1(I_1 - I_2) = 4(I_1 - I_2)$ [V] over the resistor, the difference because I_1 and I_2 flow through the resistor in opposite directions. By Kirchhoff's Voltage Law the sum of these drops equals the voltage of the battery; that is, $I_1' + 4(I_1 - I_2) = 12$, hence

$$I_1' = -4I_1 + 4I_2 + 12. \tag{4a}$$

In the right loop, the voltage drops are $R_2I_2 = 6I_2$ [V] and $R_1(I_2 - I_1) = 4(I_2 - I_1)$ [V] over the resistors and $(I/C)\int I_2\,dt = 4\int I_2\,dt$ [V] over the capacitor, and their sum is zero,

$$6I_2 + 4(I_2 - I_1) + 4\int I_2\,dt = 0 \qquad \text{or} \qquad 10I_2 - 4I_1 + 4\int I_2\,dt = 0.$$

Division by 10 and differentiation gives $I_2' - 0.4I_1' + 0.4I_2 = 0$.

To simplify the solution process, we first get rid of $0.4I_1'$, which by (4a) equals $0.4(-4I_1 + 4I_2 + 12)$. Substitution into the present ODE gives

$$I_2' = 0.4I_1' - 0.4I_2 = 0.4(-4I_1 + 4I_2 + 12) - 0.4I_2$$

and by simplification

$$I_2' = -1.6I_1 + 1.2I_2 + 4.8. \tag{4b}$$

In matrix form, (4) is (we write $\mathbf{J}$ since $\mathbf{I}$ is the unit matrix)

$$\mathbf{J}' = \mathbf{AJ} + \mathbf{g}, \qquad \text{where} \qquad \mathbf{J} = \begin{bmatrix} I_1 \\ I_2 \end{bmatrix}, \quad \mathbf{A} = \begin{bmatrix} -4.0 & 4.0 \\ -1.6 & 1.2 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} 12.0 \\ 4.8 \end{bmatrix}. \tag{5}$$

***Step 2.* Solving (5).** Because of the vector $\mathbf{g}$ this is a *nonhomogeneous* system, and we try to proceed as for a single ODE, solving first the *homogeneous* system $\mathbf{J}' = \mathbf{AJ}$ (thus $\mathbf{J}' - \mathbf{AJ} = \mathbf{0}$) by substituting $\mathbf{J} = \mathbf{x}e^{\lambda t}$. This gives

$$\mathbf{J}' = \lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{x}e^{\lambda t}, \qquad \text{hence} \qquad \mathbf{Ax} = \lambda\mathbf{x}.$$

Hence, to obtain a nontrivial solution, we again need the eigenvalues and eigenvectors. For the present matrix $\mathbf{A}$ they are derived in Example 1 in Sec. 4.0:

$$\lambda_1 = -2, \qquad \mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ 1 \end{bmatrix}; \qquad \lambda_2 = -0.8, \qquad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix}.$$

Hence a "general solution" of the homogeneous system is

$$\mathbf{J}_h = c_1\mathbf{x}^{(1)}e^{-2t} + c_2\mathbf{x}^{(2)}e^{-0.8t}.$$

For a particular solution of the nonhomogeneous system (5), since $\mathbf{g}$ is constant, we try a constant column vector $\mathbf{J}_p = \mathbf{a}$ with components a_1, a_2. Then $\mathbf{J}_p' = \mathbf{0}$, and substitution into (5) gives $\mathbf{Aa} + \mathbf{g} = \mathbf{0}$; in components,

$$\begin{aligned} -4.0a_1 + 4.0a_2 + 12.0 &= 0 \\ -1.6a_1 + 1.2a_2 + \ \ 4.8 &= 0. \end{aligned}$$

The solution is $a_1 = 3$, $a_2 = 0$; thus $\mathbf{a} = \begin{bmatrix} 3 \\ 0 \end{bmatrix}$. Hence

$$\mathbf{J} = \mathbf{J}_h + \mathbf{J}_p = c_1\mathbf{x}^{(1)}e^{-2t} + c_2\mathbf{x}^{(2)}e^{-0.8t} + \mathbf{a}; \tag{6}$$

in components,

$$\begin{aligned} I_1 &= 2c_1e^{-2t} + \quad c_2e^{-0.8t} + 3 \\ I_2 &= \ \ c_1e^{-2t} + 0.8c_2e^{-0.8t}. \end{aligned}$$

The initial conditions give

$$I_1(0) = 2c_1 + c_2 + 3 = 0$$
$$I_2(0) = c_1 + 0.8c_2 = 0.$$

Hence $c_1 = -4$ and $c_2 = 5$. As the solution of our problem we thus obtain

(7) $$\mathbf{J} = -4\mathbf{x}^{(1)}e^{-2t} + 5\mathbf{x}^{(2)}e^{-0.8t} + \mathbf{a}.$$

In components (Fig. 80b),

$$I_1 = -8e^{-2t} + 5e^{-0.8t} + 3$$
$$I_2 = -4e^{-2t} + 4e^{-0.8t}.$$

Now comes an important idea, on which we shall elaborate further, beginning in Sec. 4.3. Figure 80a shows $I_1(t)$ and $I_2(t)$ as two separate curves. Figure 80b shows these two currents as a single curve $[I_1(t), I_2(t)]$ in the I_1I_2-plane. This is a parametric representation with time t as the parameter. It is often important to know in which sense such a curve is traced. This can be indicated by an arrow in the sense of increasing t, as is shown. The I_1I_2-plane is called the **phase plane** of our system (5), and the curve in Fig. 80b is called a **trajectory**. We shall see that such "**phase plane representations**" are far more important than graphs as in Fig. 80a because they will give a much better qualitative overall impression of the general behavior of whole families of solutions, not merely of one solution as in the present case. ■

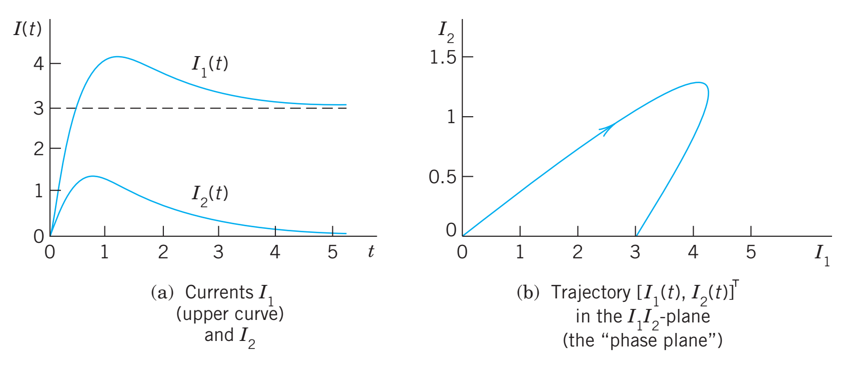


Fig. 80. Currents in Example 2

Remark. In both examples, by growing the dimension of the problem (from one tank to two tanks or one circuit to two circuits) we also increased the number of ODEs (from one ODE to two ODEs). This "growth" in the problem being reflected by an "increase" in the mathematical model is attractive and affirms the quality of our mathematical modeling and theory.

Conversion of an *n*th-Order ODE to a System

We show that an nth-order ODE of the general form (8) (see Theorem 1) can be converted to a system of n first-order ODEs. This is practically and theoretically important—practically because it permits the study and solution of single ODEs by methods for systems, and theoretically because it opens a way of including the theory of higher order ODEs into that of first-order systems. This conversion is another reason for the importance of systems, in addition to their use as models in various basic applications. The idea of the conversion is simple and straightforward, as follows.

THEOREM 1

Conversion of an ODE

An nth-order ODE

$$y^{(n)} = F(t, y, y', \cdots, y^{(n-1)}) \tag{8}$$

can be converted to a system of n first-order ODEs by setting

$$y_1 = y, \quad y_2 = y', \quad y_3 = y'', \cdots, y_n = y^{(n-1)}. \tag{9}$$

This system is of the form

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{n-1}' &= y_n \\ y_n' &= F(t, y_1, y_2, \cdots, y_n). \end{aligned} \tag{10}$$

PROOF The first $n - 1$ of these n ODEs follows immediately from (9) by differentiation. Also, $y_n' = y^{(n)}$ by (9), so that the last equation in (10) results from the given ODE (8). ■

EXAMPLE 3

Mass on a Spring

To gain confidence in the conversion method, let us apply it to an old friend of ours, modeling the free motions of a mass on a spring (see Sec. 2.4)

$$my'' + cy' + ky = 0 \qquad \text{or} \qquad y'' = -\frac{c}{m}y' - \frac{k}{m}y.$$

For this ODE (8) the system (10) is linear and homogeneous,

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -\frac{k}{m}y_1 - \frac{c}{m}y_2. \end{aligned}$$

Setting $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$, we get in matrix form

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -\dfrac{k}{m} & -\dfrac{c}{m} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}.$$

The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -\dfrac{k}{m} & -\dfrac{c}{m} - \lambda \end{vmatrix} = \lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

It agrees with that in Sec. 2.4. For an illustrative computation, let $m = 1, c = 2$, and $k = 0.75$. Then

$$\lambda^2 + 2\lambda + 0.75 = (\lambda + 0.5)(\lambda + 1.5) = 0.$$

This gives the eigenvalues $\lambda_1 = -0.5$ and $\lambda_2 = -1.5$. Eigenvectors follow from the first equation in $\mathbf{A} - \lambda\mathbf{I} = \mathbf{0}$, which is $-\lambda x_1 + x_2 = 0$. For λ_1 this gives $0.5x_1 + x_2 = 0$, say, $x_1 = 2$, $x_2 = -1$. For $\lambda_2 = -1.5$ it gives $1.5x_1 + x_2 = 0$, say, $x_1 = 1$, $x_2 = -1.5$. These eigenvectors

$$\mathbf{x}^{(1)} = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \quad \mathbf{x}^{(2)} = \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} \quad \text{give} \quad \mathbf{y} = c_1 \begin{bmatrix} 2 \\ -1 \end{bmatrix} e^{-0.5t} + c_2 \begin{bmatrix} 1 \\ -1.5 \end{bmatrix} e^{-1.5t}.$$

This vector solution has the first component

$$y = y_1 = 2c_1e^{-0.5t} + c_2e^{-1.5t}$$

which is the expected solution. The second component is its derivative

$$y_2 = y_1' = y' = -c_1e^{-0.5t} - 1.5c_2e^{-1.5t}.$$ ■

PROBLEM SET 4.1

1–6 MIXING PROBLEMS

1. Find out, without calculation, whether doubling the flow rate in Example 1 has the same effect as halfing the tank sizes. (Give a reason.)
2. What happens in Example 1 if we replace T_1 by a tank containing 200 gal of water and 150 lb of fertilizer dissolved in it?
3. Derive the eigenvectors in Example 1 without consulting this book.
4. In Example 1 find a "general solution" for any ratio $a =$ (flow rate)/(tank size), tank sizes being equal. Comment on the result.
5. If you extend Example 1 by a tank T_3 of the same size as the others and connected to T_2 by two tubes with flow rates as between T_1 and T_2, what system of ODEs will you get?
6. Find a "general solution" of the system in Prob. 5.

7–9 ELECTRICAL NETWORK

In Example 2 find the currents:

7. If the initial currents are 0 A and -3 A (minus meaning that $I_2(0)$ flows against the direction of the arrow).
8. If the capacitance is changed to $C = 5/27$ F. (General solution only.)
9. If the initial currents in Example 2 are 28 A and 14 A.

10–13 CONVERSION TO SYSTEMS

Find a general solution of the given ODE **(a)** by first converting it to a system, **(b)**, as given. Show the details of your work.

10. $y'' + 3y' + 2y = 0$
11. $4y'' - 15y' - 4y = 0$
12. $y''' + 2y'' - y' - 2y = 0$
13. $y'' + 2y' - 24y = 0$

14. **TEAM PROJECT. Two Masses on Springs. (a)** Set up the model for the (undamped) system in Fig. 81. **(b)** Solve the system of ODEs obtained. *Hint.* Try $\mathbf{y} = \mathbf{x}e^{\omega t}$ and set $\omega^2 = \lambda$. Proceed as in Example 1 or 2. **(c)** Describe the influence of initial conditions on the possible kind of motions.

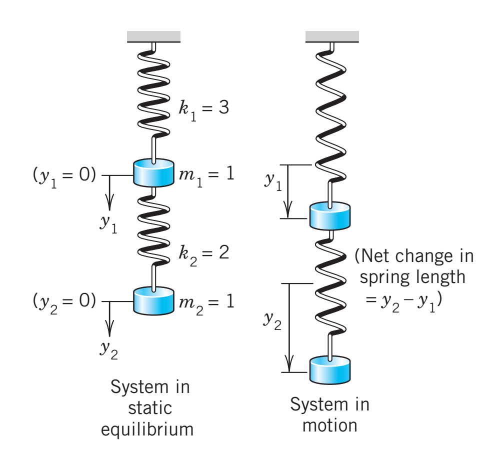


Fig. 81. Mechanical system in Team Project

15. **CAS EXPERIMENT. Electrical Network. (a)** In Example 2 choose a sequence of values of C that increases beyond bound, and compare the corresponding sequences of eigenvalues of $\mathbf{A}$. What limits of these sequences do your numeric values (approximately) suggest?

    (b) Find these limits analytically.

    (c) Explain your result physically.

    (d) Below what value (approximately) must you decrease C to get vibrations?

4.2 Basic Theory of Systems of ODEs. Wronskian

In this section we discuss some basic concepts and facts about system of ODEs that are quite similar to those for single ODEs.

The first-order systems in the last section were special cases of the more general system

$$\text{(1)}\qquad \begin{aligned} y_1' &= f_1(t, y_1, \cdots, y_n) \\ y_2' &= f_2(t, y_1, \cdots, y_n) \\ &\cdots \\ y_n' &= f_n(t, y_1, \cdots, y_n). \end{aligned}$$

We can write the system (1) as a vector equation by introducing the column vectors $\mathbf{y} = [y_1 \ \cdots \ y_n]^\mathsf{T}$ and $\mathbf{f} = [f_1 \ \cdots \ f_n]^\mathsf{T}$ (where T means *transposition* and saves us the space that would be needed for writing $\mathbf{y}$ and $\mathbf{f}$ as columns). This gives

$$\text{(1)}\qquad \mathbf{y}' = \mathbf{f}(t, \mathbf{y}).$$

This system (1) includes almost all cases of practical interest. For $n = 1$ it becomes $y_1' = f_1(t, y_1)$ or, simply, $y' = f(t, y)$, well known to us from Chap. 1.

A **solution** of (1) on some interval $a < t < b$ is a set of n differentiable functions

$$y_1 = h_1(t), \quad \cdots, \quad y_n = h_n(t)$$

on $a < t < b$ that satisfy (1) throughout this interval. In vector from, introducing the "*solution vector*" $\mathbf{h} = [h_1 \ \cdots \ h_n]^\mathsf{T}$ (a column vector!) we can write

$$\mathbf{y} = \mathbf{h}(t).$$

An **initial value problem** for (1) consists of (1) and n given **initial conditions**

$$\text{(2)}\qquad y_1(t_0) = K_1, \quad y_2(t_0) = K_2, \quad \cdots, \quad y_n(t_0) = K_n,$$

in vector form, $\mathbf{y}(t_0) = \mathbf{K}$, where t_0 is a specified value of t in the interval considered and the components of $\mathbf{K} = [K_1 \ \cdots \ K_n]^\mathsf{T}$ are given numbers. Sufficient conditions for the existence and uniqueness of a solution of an initial value problem (1), (2) are stated in the following theorem, which extends the theorems in Sec. 1.7 for a single equation. (For a proof, see Ref. [A7].)

THEOREM 1

Existence and Uniqueness Theorem

Let $f_1, \cdots, f_n$ in (1) *be continuous functions having continuous partial derivatives $\partial f_1/\partial y_1, \cdots, \partial f_1/\partial y_n, \cdots, \partial f_n/\partial y_n$ in some domain R of $ty_1y_2\cdots y_n$-space containing the point $(t_0, K_1, \cdots, K_n)$. Then* (1) *has a solution on some interval $t_0 - \alpha < t < t_0 + \alpha$ satisfying* (2), *and this solution is unique.*

Linear Systems

Extending the notion of a *linear* ODE, we call (1) a **linear system** if it is linear in $y_1, \cdots, y_n$; that is, if it can be written

(3)
$$\begin{aligned} y_1' &= a_{11}(t)y_1 + \cdots + a_{1n}(t)y_n + g_1(t) \\ &\vdots \\ y_n' &= a_{n1}(t)y_1 + \cdots + a_{nn}(t)y_n + g_n(t). \end{aligned}$$

As a vector equation this becomes

(3)
$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}$$

where
$$\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ . & \cdots & . \\ a_{n1} & \cdots & a_{nn} \end{bmatrix}, \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \qquad \mathbf{g} = \begin{bmatrix} g_1 \\ \vdots \\ g_n \end{bmatrix}.$$

This system is called **homogeneous** if $\mathbf{g} = \mathbf{0}$, so that it is

(4)
$$\mathbf{y}' = \mathbf{A}\mathbf{y}.$$

If $\mathbf{g} \neq \mathbf{0}$, then (3) is called **nonhomogeneous**. For example, the systems in Examples 1 and 3 of Sec. 4.1 are homogeneous. The system in Example 2 of that section is nonhomogeneous.

For a linear system (3) we have $\partial f_1/\partial y_1 = a_{11}(t), \cdots, \partial f_n/\partial y_n = a_{nn}(t)$ in Theorem 1. Hence for a linear system we simply obtain the following.

THEOREM 2

Existence and Uniqueness in the Linear Case

Let the a_{jk}'s and g_j's in (3) *be continuous functions of t on an open interval $\alpha < t < \beta$ containing the point $t = t_0$. Then* (3) *has a solution $\mathbf{y}(t)$ on this interval satisfying* (2), *and this solution is unique.*

As for a single homogeneous linear ODE we have

THEOREM 3

Superposition Principle or Linearity Principle

*If $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are solutions of the **homogeneous linear** system* (4) *on some interval, so is any linear combination $\mathbf{y} = c_1\mathbf{y}^{(1)} + c_1\mathbf{y}^{(2)}$.*

PROOF Differentiating and using (4), we obtain

$$\begin{aligned} \mathbf{y}' &= [c_1\mathbf{y}^{(1)} + c_1\mathbf{y}^{(2)}]' \\ &= c_1\mathbf{y}^{(1)\prime} + c_2\mathbf{y}^{(2)\prime} \\ &= c_1\mathbf{A}\mathbf{y}^{(1)} + c_2\mathbf{A}\mathbf{y}^{(2)} \\ &= \mathbf{A}(c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)}) = \mathbf{A}\mathbf{y}. \end{aligned}$$

■

The general theory of linear systems of ODEs is quite similar to that of a single linear ODE in Secs. 2.6 and 2.7. To see this, we explain the most basic concepts and facts. For proofs we refer to more advanced texts, such as [A7].

Basis. General Solution. Wronskian

By a **basis** or a **fundamental system** of solutions of the homogeneous system (4) on some interval J we mean a linearly independent set of n solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ of (4) on that interval. (We write J because we need $\mathbf{I}$ to denote the unit matrix.) We call a corresponding linear combination

$$\mathbf{y} = c_1\mathbf{y}^{(1)} \cdots + c_n\mathbf{y}^{(n)} \qquad (c_1, \cdots, c_n \text{ arbitrary}) \tag{5}$$

a **general solution** of (4) on J. It can be shown that if the $a_{jk}(t)$ in (4) are continuous on J, then (4) has a basis of solutions on J, hence a general solution, which includes every solution of (4) on J.

We can write n solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ of (4) on some interval J as columns of an $n \times n$ matrix

$$\mathbf{Y} = [\mathbf{y}^{(1)} \quad \cdots \quad \mathbf{y}^{(n)}]. \tag{6}$$

The determinant of $\mathbf{Y}$ is called the **Wronskian** of $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$, written

$$W(\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}) = \begin{vmatrix} y_1^{(1)} & y_1^{(2)} & \cdots & y_1^{(n)} \\ y_2^{(1)} & y_2^{(2)} & \cdots & y_2^{(n)} \\ \cdot & \cdot & \cdots & \cdot \\ y_n^{(1)} & y_n^{(2)} & \cdots & y_n^{(n)} \end{vmatrix}. \tag{7}$$

The columns are these solutions, each in terms of components. These solutions form a basis on J if and only if W is not zero at any t_1 in this interval. W is either identically zero or nowhere zero in J. (This is similar to Secs. 2.6 and 3.1.)

If the solutions $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ in (5) form a basis (a fundamental system), then (6) is often called a **fundamental matrix**. Introducing a column vector $\mathbf{c} = [c_1 \quad c_2 \quad \cdots \quad c_n]^{\mathsf{T}}$, we can now write (5) simply as

$$\mathbf{y} = \mathbf{Y}\mathbf{c}. \tag{8}$$

Furthermore, we can relate (7) to Sec. 2.6, as follows. If y and z are solutions of a second-order homogeneous linear ODE, their Wronskian is

$$W(y, z) = \begin{vmatrix} y & z \\ y' & z' \end{vmatrix}.$$

To write this ODE as a system, we have to set $y = y_1$, $y' = y_1' = y_2$ and similarly for z (see Sec. 4.1). But then $W(y, z)$ becomes (7), except for notation.

4.3 Constant-Coefficient Systems. Phase Plane Method

Continuing, we now assume that our homogeneous linear system

$$\mathbf{y}' = \mathbf{A}\mathbf{y} \tag{1}$$

under discussion has ***constant coefficients***, so that the $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has entries not depending on t. We want to solve (1). Now a single ODE $y' = ky$ has the solution $y = Ce^{kt}$. So let us try

$$\mathbf{y} = \mathbf{x}e^{\lambda t}. \tag{2}$$

Substitution into (1) gives $\mathbf{y}' = \lambda \mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{x}e^{\lambda t}$. Dividing by $e^{\lambda t}$, we obtain the **eigenvalue problem**

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x}. \tag{3}$$

Thus the nontrivial solutions of (1) (solutions that are not zero vectors) are of the form (2), where λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ is a corresponding eigenvector.

We assume that $\mathbf{A}$ has a linearly independent set of n eigenvectors. This holds in most applications, in particular if $\mathbf{A}$ is symmetric ($a_{kj} = a_{jk}$) or skew-symmetric ($a_{kj} = -a_{jk}$) or has n *different* eigenvalues.

Let those eigenvectors be $\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(n)}$ and let them correspond to eigenvalues $\lambda_1, \cdots, \lambda_n$ (which may be all different, or some—or even all—may be equal). Then the corresponding solutions (2) are

$$\mathbf{y}^{(4)} = \mathbf{x}^{(1)}e^{\lambda_1 t}, \quad \cdots, \quad \mathbf{y}^{(n)} = \mathbf{x}^{(n)}e^{\lambda_n t}. \tag{4}$$

Their Wronskian $W = W(\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)})$ [(7) in Sec. 4.2] is given by

$$W = (\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}) = \begin{vmatrix} x_1^{(1)}e^{\lambda_1 t} & \cdots & x_1^{(n)}e^{\lambda_n t} \\ x_2^{(1)}e^{\lambda_1 t} & \cdots & x_2^{(n)}e^{\lambda_n t} \\ \cdot & \cdots & \cdot \\ x_n^{(1)}e^{\lambda_1 t} & \cdots & x_n^{(n)}e^{\lambda_n t} \end{vmatrix} = e^{\lambda_1 t + \cdots + \lambda_n t} \begin{vmatrix} x_1^{(1)} & \cdots & x_1^{(n)} \\ x_2^{(1)} & \cdots & x_2^{(n)} \\ \cdot & \cdots & \cdot \\ x_n^{(1)} & \cdots & x_n^{(n)} \end{vmatrix}.$$

On the right, the exponential function is never zero, and the determinant is not zero either because its columns are the n linearly independent eigenvectors. This proves the following theorem, whose assumption is true if the matrix $\mathbf{A}$ is symmetric or skew-symmetric, or if the n eigenvalues of $\mathbf{A}$ are all different.

THEOREM 1

General Solution

If the constant matrix $\mathbf{A}$ *in the system* (1) *has a linearly independent set of* n *eigenvectors, then the corresponding solutions* $\mathbf{y}^{(1)}, \cdots, \mathbf{y}^{(n)}$ *in* (4) *form a basis of solutions of* (1), *and the corresponding general solution is*

$$\mathbf{y} = c_1\mathbf{x}^{(1)}e^{\lambda_1 t} + \cdots + c_n\mathbf{x}^{(n)}e^{\lambda_n t}. \tag{5}$$

How to Graph Solutions in the Phase Plane

We shall now concentrate on systems (1) with constant coefficients consisting of two ODEs

$$\mathbf{y}' = \mathbf{A}\mathbf{y}; \qquad \text{in components,} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned} \tag{6}$$

Of course, we can graph solutions of (6),

$$\mathbf{y}(t) = \begin{bmatrix} y_1(t) \\ y_2(t) \end{bmatrix}, \tag{7}$$

as two curves over the t-axis, one for each component of $\mathbf{y}(t)$. (Figure 80a in Sec. 4.1 shows an example.) But we can also graph (7) as a single curve in the y_1y_2-plane. This is a *parametric representation* (*parametric equation*) with parameter t. (See Fig. 80b for an example. Many more follow. Parametric equations also occur in calculus.) Such a curve is called a **trajectory** (or sometimes an *orbit* or *path*) of (6). The y_1y_2-plane is called the **phase plane**.[1] If we fill the phase plane with trajectories of (6), we obtain the so-called **phase portrait** of (6).

Studies of solutions in the phase plane have become quite important, along with advances in computer graphics, because a phase portrait gives a good general qualitative impression of the entire family of solutions. Consider the following example, in which we develop such a phase portrait.

EXAMPLE 1 **Trajectories in the Phase Plane (Phase Portrait)**

Find and graph solutions of the system.

In order to see what is going on, let us find and graph solutions of the system

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}\mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= -3y_1 + y_2 \\ y_2' &= y_1 - 3y_2. \end{aligned} \tag{8}$$

---

[1] A name that comes from physics, where it is the y-(mv)-plane, used to plot a motion in terms of position y and velocity $y' = v$ (m = mass); but the name is now used quite generally for the y_1y_2-plane.

The use of the phase plane is a **qualitative method**, a method of obtaining general qualitative information on solutions without actually solving an ODE or a system. This method was created by HENRI POINCARÉ (1854–1912), a great French mathematician, whose work was also fundamental in complex analysis, divergent series, topology, and astronomy.

Solution. By substituting $\mathbf{y} = \mathbf{x}e^{\lambda t}$ and $\mathbf{y}' = \lambda\mathbf{x}e^{\lambda t}$ and dropping the exponential function we get $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. The characteristic equation is

$$\det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -3-\lambda & 1 \\ 1 & -3-\lambda \end{vmatrix} = \lambda^2 + 6\lambda + 8 = 0.$$

This gives the eigenvalues $\lambda_1 = -2$ and $\lambda_2 = -4$. Eigenvectors are then obtained from

$$(-3-\lambda)x_1 + x_2 = 0.$$

For $\lambda_1 = -2$ this is $-x_1 + x_2 = 0$. Hence we can take $\mathbf{x}^{(1)} = [1 \quad 1]^\mathsf{T}$. For $\lambda_2 = -4$ this becomes $x_1 + x_2 = 0$, and an eigenvector is $\mathbf{x}^{(2)} = [1 \quad -1]^\mathsf{T}$. This gives the general solution

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t}.$$

Figure 82 shows a phase portrait of some of the trajectories (to which more trajectories could be added if so desired). The two straight trajectories correspond to $c_1 = 0$ and $c_2 = 0$ and the others to other choices of c_1, c_2. ■

The method of the phase plane is particularly valuable in the frequent cases when solving an ODE or a system is inconvenient of impossible.

Critical Points of the System (6)

The point $\mathbf{y} = \mathbf{0}$ in Fig. 82 seems to be a common point of all trajectories, and we want to explore the reason for this remarkable observation. The answer will follow by calculus. Indeed, from (6) we obtain

$$\frac{dy_2}{dy_1} = \frac{y_2'\,dt}{y_1'\,dt} = \frac{y_2'}{y_1'} = \frac{a_{21}y_1 + a_{22}y_2}{a_{11}y_1 + a_{12}y_2}. \tag{9}$$

This associates with every point $P: (y_1, y_2)$ a unique tangent direction dy_2/dy_1 of the trajectory passing through P, except for the point $P = P_0: (0, 0)$, where the right side of (9) becomes $0/0$. This point P_0, at which dy_2/dy_1 becomes undetermined, is called a **critical point** of (6).

Five Types of Critical Points

There are five types of critical points depending on the geometric shape of the trajectories near them. They are called **improper nodes, proper nodes, saddle points, centers**, and **spiral points**. We define and illustrate them in Examples 1–5.

EXAMPLE 1 (*Continued*) Improper Node (Fig. 82)

An **improper node** is a critical point P_0 at which all the trajectories, except for two of them, have the same limiting direction of the tangent. The two exceptional trajectories also have a limiting direction of the tangent at P_0 which, however, is different.

The system (8) has an improper node at $\mathbf{0}$, as its phase portrait Fig. 82 shows. The common limiting direction at $\mathbf{0}$ is that of the eigenvector $\mathbf{x}^{(1)} = [1 \quad 1]^\mathsf{T}$ because e^{-4t} goes to zero faster than e^{-2t} as t increases. The two exceptional limiting tangent directions are those of $\mathbf{x}^{(2)} = [1 \quad -1]^\mathsf{T}$ and $-\mathbf{x}^{(2)} = [-1 \quad 1]^\mathsf{T}$. ■

EXAMPLE 2 **Proper Node (Fig. 83)**

A **proper node** is a critical point P_0 at which every trajectory has a definite limiting direction and for any given direction $\mathbf{d}$ at P_0 there is a trajectory having $\mathbf{d}$ as its limiting direction.

The system

$$(10) \qquad \mathbf{y}' = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_1 \\ y_2' &= y_2 \end{aligned}$$

has a proper node at the origin (see Fig. 83). Indeed, the matrix is the unit matrix. Its characteristic equation $(1 - \lambda)^2 = 0$ has the root $\lambda = 1$. Any $\mathbf{x} \neq \mathbf{0}$ is an eigenvector, and we can take $[1 \quad 0]^{\mathsf{T}}$ and $[0 \quad 1]^{\mathsf{T}}$. Hence a general solution is

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^t \qquad \text{or} \qquad \begin{aligned} y_1 &= c_1 e^t \\ y_2 &= c_2 e^t \end{aligned} \qquad \text{or} \qquad c_1 y_2 = c_2 y_1.$$

■

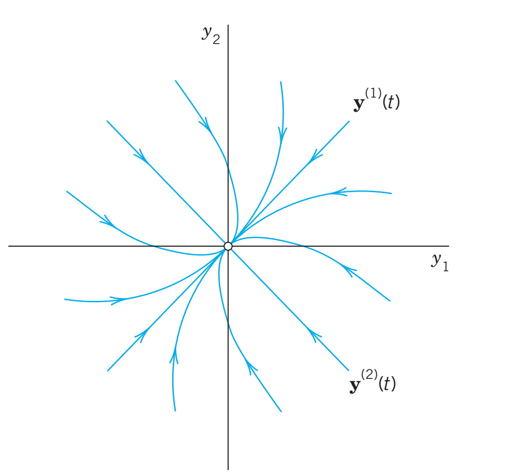


Fig. 82. Trajectories of the system (8) (Improper node)

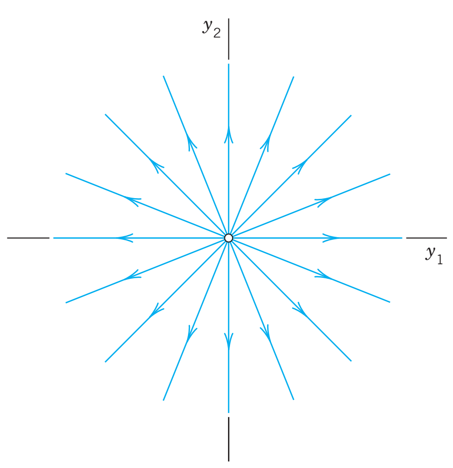


Fig. 83. Trajectories of the system (10) (Proper node)

EXAMPLE 3 **Saddle Point (Fig. 84)**

A **saddle point** is a critical point P_0 at which there are two incoming trajectories, two outgoing trajectories, and all the other trajectories in a neighborhood of P_0 bypass P_0.

The system

$$(11) \qquad \mathbf{y}' = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= \ \ y_1 \\ y_1' &= -y_2 \end{aligned}$$

has a saddle point at the origin. Its characteristic equation $(1 - \lambda)(-1 - \lambda) = 0$ has the roots $\lambda_1 = 1$ and $\lambda_2 = -1$. For $\lambda = 1$ an eigenvector $[1 \quad 0]^{\mathsf{T}}$ is obtained from the second row of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, that is, $0x_1 + (-1 - 1)x_2 = 0$. For $\lambda_2 = -1$ the first row gives $[0 \quad 1]^{\mathsf{T}}$. Hence a general solution is

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^t + c_2 \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{-t} \qquad \text{or} \qquad \begin{aligned} y_1 &= c_1 e^t \\ y_2 &= c_2 e^{-t} \end{aligned} \qquad \text{or} \qquad y_1 y_2 = \text{const.}$$

This is a family of hyperbolas (and the coordinate axes); see Fig. 84. ■

EXAMPLE 4 **Center (Fig. 85)**

A **center** is a critical point that is enclosed by infinitely many closed trajectories.

The system

$$\text{(12)} \qquad \mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -4 & 0 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{array}{ll} \text{(a)} & y_1' = y_2 \\ \text{(b)} & y_2' = -4y_1 \end{array}$$

has a center at the origin. The characteristic equation $\lambda^2 + 4 = 0$ gives the eigenvalues $2i$ and $-2i$. For $2i$ an eigenvector follows from the first equation $-2ix_1 + x_2 = 0$ of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, say, $[1 \quad 2i]^\mathsf{T}$. For $\lambda = -2i$ that equation is $-(-2i)x_1 + x_2 = 0$ and gives, say, $[1 \quad -2i]^\mathsf{T}$. Hence a complex general solution is

$$\text{(12*)} \qquad \mathbf{y} = c_1 \begin{bmatrix} 1 \\ 2i \end{bmatrix} e^{2it} + c_2 \begin{bmatrix} 1 \\ -2i \end{bmatrix} e^{-2it}, \qquad \text{thus} \qquad \begin{aligned} y_1 &= c_1 e^{2it} + c_2 e^{-2it} \\ y_2 &= 2ic_1 e^{2it} - 2ic_2 e^{-2it}. \end{aligned}$$

A real solution is obtained from (12*) by the Euler formula or directly from (12) by a trick. (Remember the trick and call it a method when you apply it again.) Namely, the left side of (a) times the right side of (b) is $-4y_1y_1'$. This must equal the left side of (b) times the right side of (a). Thus,

$$-4y_1y_1' = y_2y_2'. \qquad \text{By integration,} \qquad 2y_1^2 + \tfrac{1}{2}y_2^2 = \text{const.}$$

This is a family of ellipses (see Fig. 85) enclosing the center at the origin. ■

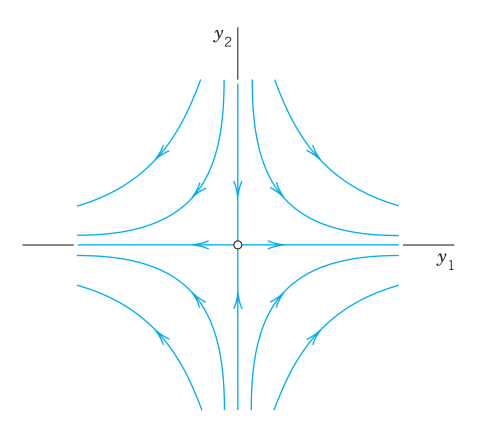


Fig. 84. Trajectories of the system (11) (Saddle point)

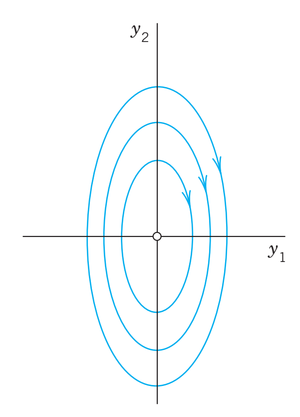


Fig. 85. Trajectories of the system (12) (Center)

EXAMPLE 5 **Spiral Point (Fig. 86)**

A **spiral point** is a critical point P_0 about which the trajectories spiral, approaching P_0 as $t \to \infty$ (or tracing these spirals in the opposite sense, away from P_0).

The system

$$\text{(13)} \qquad \mathbf{y}' = \begin{bmatrix} -1 & 1 \\ -1 & -1 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= -y_1 + y_2 \\ y_2' &= -y_1 - y_2 \end{aligned}$$

has a spiral point at the origin, as we shall see. The characteristic equation is $\lambda^2 + 2\lambda + 2 = 0$. It gives the eigenvalues $-1 + i$ and $-1 - i$. Corresponding eigenvectors are obtained from $(-1 - \lambda)x_1 + x_2 = 0$. For

$\lambda = -1 + i$ this becomes $-ix_1 + x_2 = 0$ and we can take $[1 \quad i]^T$ as an eigenvector. Similarly, an eigenvector corresponding to $-1 - i$ is $[1 \quad -i]^T$. This gives the complex general solution

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ i \end{bmatrix} e^{(-1+i)t} + c_2 \begin{bmatrix} 1 \\ -i \end{bmatrix} e^{(-1-i)t}.$$

The next step would be the transformation of this complex solution to a real general solution by the Euler formula. But, as in the last example, we just wanted to see what eigenvalues to expect in the case of a spiral point. Accordingly, we start again from the beginning and instead of that rather lengthy systematic calculation we use a shortcut. We multiply the first equation in (13) by y_1, the second by y_2, and add, obtaining

$$y_1 y_1' + y_2 y_2' = -(y_1^2 + y_2^2).$$

We now introduce polar coordinates r, t, where $r^2 = y_1^2 + y_2^2$. Differentiating this with respect to t gives $2rr' = 2y_1 y_1' + 2y_2 y_2'$. Hence the previous equation can be written

$$rr' = -r^2, \qquad \text{Thus,} \qquad r' = -r, \qquad dr/r = -dt, \qquad \ln |r| = -t + c^*, \qquad r = ce^{-t}.$$

For each real c this is a spiral, as claimed (see Fig. 86). ■

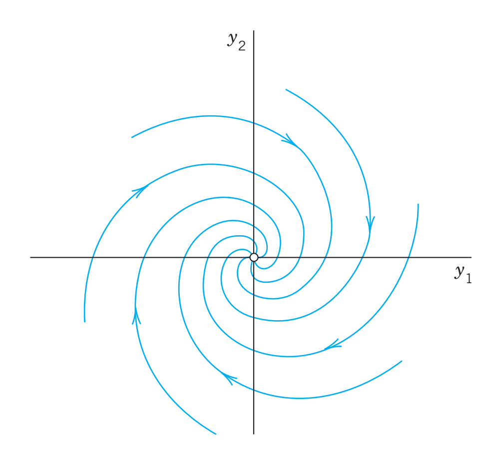


Fig. 86. Trajectories of the system (13) (Spiral point)

EXAMPLE 6 No Basis of Eigenvectors Available. Degenerate Node (Fig. 87)

This cannot happen if $\mathbf{A}$ in (1) is symmetric ($a_{kj} = a_{jk}$, as in Examples 1–3) or skew-symmetric ($a_{kj} = -a_{jk}$, thus $a_{jj} = 0$). And it does not happen in many other cases (see Examples 4 and 5). Hence it suffices to explain the method to be used by an example.

Find and graph a general solution of

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 4 & 1 \\ -1 & 2 \end{bmatrix} \mathbf{y}. \tag{14}$$

Solution. $\mathbf{A}$ is not skew-symmetric! Its characteristic equation is

$$\det(\mathbf{A} - \lambda \mathbf{I}) = \begin{vmatrix} 4-\lambda & 1 \\ -1 & 2-\lambda \end{vmatrix} = \lambda^2 - 6\lambda + 9 = (\lambda - 3)^2 = 0.$$

It has a double root $\lambda = 3$. Hence eigenvectors are obtained from $(4 - \lambda)x_1 + x_2 = 0$, thus from $x_1 + x_2 = 0$, say, $\mathbf{x}^{(1)} = [1 \quad -1]^\mathsf{T}$ and nonzero multiples of it (which do not help). The method now is to substitute

$$\mathbf{y}^{(2)} = \mathbf{x}te^{\lambda t} + \mathbf{u}e^{\lambda t}$$

with constant $\mathbf{u} = [u_1 \quad u_2]^\mathsf{T}$ into (14). (The $\mathbf{x}t$-term alone, the analog of what we did in Sec. 2.2 in the case of a double root, would not be enough. Try it.) This gives

$$\mathbf{y}^{(2)\prime} = \mathbf{x}e^{\lambda t} + \lambda\mathbf{x}te^{\lambda t} + \lambda\mathbf{u}e^{\lambda t} = \mathbf{A}\mathbf{y}^{(2)} = \mathbf{A}\mathbf{x}te^{\lambda t} + \mathbf{A}\mathbf{u}e^{\lambda t}.$$

On the right, $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Hence the terms $\lambda\mathbf{x}te^{\lambda t}$ cancel, and then division by $e^{\lambda t}$ gives

$$\mathbf{x} + \lambda\mathbf{u} = \mathbf{A}\mathbf{u}, \qquad \text{thus} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = \mathbf{x}.$$

Here $\lambda = 3$ and $\mathbf{x} = [1 \quad -1]^\mathsf{T}$, so that

$$(\mathbf{A} - 3\mathbf{I})\mathbf{u} = \begin{bmatrix} 4-3 & 1 \\ -1 & 2-3 \end{bmatrix}\mathbf{u} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \text{thus} \qquad \begin{aligned} u_1 + u_2 &= 1 \\ -u_1 - u_2 &= -1. \end{aligned}$$

A solution, linearly independent of $\mathbf{x} = [1 \quad -1]^\mathsf{T}$, is $\mathbf{u} = [0 \quad 1]^\mathsf{T}$. This yields the answer (Fig. 87)

$$\mathbf{y} = c_1\mathbf{y}^{(1)} + c_2\mathbf{y}^{(2)} = c_1\begin{bmatrix} 1 \\ -1 \end{bmatrix}e^{3t} + c_2\left(\begin{bmatrix} 1 \\ -1 \end{bmatrix}t + \begin{bmatrix} 0 \\ 1 \end{bmatrix}\right)e^{3t}.$$

The critical point at the origin is often called a **degenerate node**. $c_1\mathbf{y}^{(1)}$ gives the heavy straight line, with $c_1 > 0$ the lower part and $c_1 < 0$ the upper part of it. $\mathbf{y}^{(2)}$ gives the right part of the heavy curve from 0 through the second, first, and—finally—fourth quadrants. $-\mathbf{y}^{(2)}$ gives the other part of that curve. ■

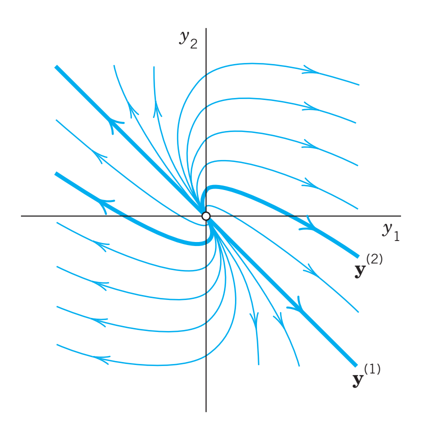


Fig. 87. Degenerate node in Example 6

We mention that for a system (1) with three or more equations and a triple eigenvalue with only one linearly independent eigenvector, one will get two solutions, as just discussed, and a third linearly independent one from

$$\mathbf{y}^{(3)} = \tfrac{1}{2}\mathbf{x}t^2e^{\lambda t} + \mathbf{u}te^{\lambda t} + \mathbf{v}e^{\lambda t} \qquad \text{with } \mathbf{v} \text{ from} \qquad \mathbf{u} + \lambda\mathbf{v} = \mathbf{A}\mathbf{v}.$$

PROBLEM SET 4.3

1–9 GENERAL SOLUTION

Find a real general solution of the following systems. Show the details.

1. $y_1' = y_1 + y_2$
$y_2' = 3y_1 - y_2$

2. $y_1' = 6y_1 + 9y_2$
$y_2' = y_1 + 6y_2$

3. $y_1' = y_1 + 2y_2$
$y_2' = \frac{1}{2}y_1 + y_2$

4. $y_1' = -8y_1 - 2y_2$
$y_2' = 2y_1 - 4y_2$

5. $y_1' = 2y_1 + 5y_2$
$y_2' = 5y_1 + 12.5y_2$

6. $y_1' = 2y_1 - 2y_2$
$y_2' = 2y_1 + 2y_2$

7. $y_1' = y_2$
$y_2' = -y_1 + y_3$
$y_3' = -y_2$

8. $y_1' = 8y_1 - y_2$
$y_2' = y_1 + 10y_2$

9. $y_1' = 10y_1 - 10y_2 - 4y_3$
$y_2' = -10y_1 + y_2 - 14y_3$
$y_3' = -4y_1 - 14y_2 - 2y_3$

10–15 IVPs

Solve the following initial value problems.

10. $y_1' = 2y_1 + 2y_2$
$y_2' = 5y_1 - y_2$
$y_1(0) = 0, \quad y_2(0) = 7$

11. $y_1' = 2y_1 + 5y_2$
$y_2' = -\frac{1}{2}y_1 - \frac{3}{2}y_2$
$y_1(0) = -12, \quad y_2(0) = 0$

12. $y_1' = y_1 + 3y_2$
$y_2' = \frac{1}{3}y_1 + y_2$
$y_1(0) = 12, \quad y_2(0) = 2$

13. $y_1' = y_2$
$y_2' = y_1$
$y_1(0) = 0, \quad y_2(0) = 2$

14. $y_1' = -y_1 - y_2$
$y_2' = y_1 - y_2$
$y_1(0) = 1, \quad y_2(0) = 0$

15. $y_1' = 3y_1 + 2y_2$
$y_2' = 2y_1 + 3y_2$
$y_1(0) = 0.5, \quad y_2(0) = -0.5$

16–17 CONVERSION

Find a general solution by conversion to a single ODE.

16. The system in Prob. 8.

17. The system in Example 5 of the text.

18. Mixing problem, Fig. 88. Each of the two tanks contains 200 gal of water, in which initially 100 lb (Tank T_1) and 200 lb (Tank T_2) of fertilizer are dissolved. The inflow, circulation, and outflow are shown in Fig. 88. The mixture is kept uniform by stirring. Find the fertilizer contents $y_1(t)$ in T_1 and $y_2(t)$ in T_2.

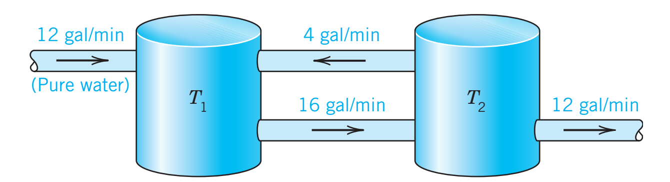


Fig. 88. Tanks in Problem 18

19. Network. Show that a model for the currents $I_1(t)$ and $I_2(t)$ in Fig. 89 is

$$\frac{1}{C}\int I_1\,dt + R(I_1 - I_2) = 0, \quad LI_2' + R(I_2 - I_1) = 0.$$

Find a general solution, assuming that $R = 3\ \Omega$, $L = 4$ H, $C = 1/12$ F.

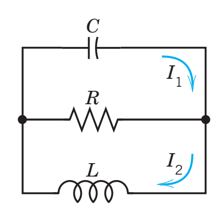


Fig. 89. Network in Problem 19

20. CAS PROJECT. Phase Portraits. Graph some of the figures in this section, in particular Fig. 87 on the degenerate node, in which the vector $\mathbf{y}^{(2)}$ depends on t. In each figure highlight a trajectory that satisfies an initial condition of your choice.

4.4 Criteria for Critical Points. Stability

We continue our discussion of homogeneous linear systems with constant coefficients (1). Let us review where we are. From Sec. 4.3 we have

$$(1) \qquad \mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \mathbf{y}, \qquad \text{in components,} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned}$$

From the examples in the last section, we have seen that we can obtain an overview of families of solution curves if we represent them parametrically as $\mathbf{y}(t) = [y_1(t) \quad y_2(t)]^\mathsf{T}$ and graph them as curves in the y_1y_2-plane, called the **phase plane**. Such a curve is called a **trajectory** of (1), and their totality is known as the **phase portrait** of (1).

Now we have seen that solutions are of the form

$$\mathbf{y}(t) = \mathbf{x}e^{\lambda t}. \qquad \text{Substitution into (1) gives} \qquad \mathbf{y}'(t) = \lambda\mathbf{x}e^{\lambda t} = \mathbf{A}\mathbf{y} = \mathbf{A}\mathbf{x}e^{\lambda t}.$$

Dropping the common factor $e^{\lambda t}$, we have

$$\textbf{(2)} \qquad \mathbf{A}\mathbf{x} = \lambda\mathbf{x}.$$

Hence $\mathbf{y}(t)$ is a (nonzero) solution of (1) if λ is an eigenvalue of $\mathbf{A}$ and $\mathbf{x}$ a corresponding eigenvector.

Our examples in the last section show that the general form of the phase portrait is determined to a large extent by the type of **critical point** of the system (1) defined as a point at which dy_2/dy_1 becomes undetermined, $0/0$; here [see (9) in Sec. 4.3]

$$\textbf{(3)} \qquad \frac{dy_2}{dy_1} = \frac{y_2'\,dt}{y_1'\,dt} = \frac{a_{21}y_1 + a_{22}y_2}{a_{11}y_1 + a_{12}y_2}.$$

We also recall from Sec. 4.3 that there are various types of critical points.

What is now new, is that we shall see how these types of critical points are related to the eigenvalues. The latter are solutions $\lambda = \lambda_1$ and λ_2 of the characteristic equation

$$(4) \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = \lambda^2 - (a_{11} + a_{22})\lambda + \det\mathbf{A} = 0.$$

This is a quadratic equation $\lambda^2 - p\lambda + q = 0$ with coefficients p, q and discriminant Δ given by

$$(5) \qquad p = a_{11} + a_{22}, \qquad q = \det\mathbf{A} = a_{11}a_{22} - a_{12}a_{21}, \qquad \Delta = p^2 - 4q.$$

From algebra we know that the solutions of this equation are

$$(6) \qquad \lambda_1 = \tfrac{1}{2}(p + \sqrt{\Delta}), \qquad \lambda_2 = \tfrac{1}{2}(p - \sqrt{\Delta}).$$

Furthermore, the product representation of the equation gives

$$\lambda^2 - p\lambda + q = (\lambda - \lambda_1)(\lambda - \lambda_2) = \lambda^2 - (\lambda_1 + \lambda_2)\lambda + \lambda_1\lambda_2.$$

Hence p is the sum and q the product of the eigenvalues. Also $\lambda_1 - \lambda_2 = \sqrt{\Delta}$ from (6). Together,

$$p = \lambda_1 + \lambda_2, \qquad q = \lambda_1\lambda_2, \qquad \Delta = (\lambda_1 - \lambda_2)^2. \tag{7}$$

This gives the criteria in Table 4.1 for classifying critical points. A derivation will be indicated later in this section.

Table 4.1 Eigenvalue Criteria for Critical Points (Derivation after Table 4.2)

Name	$p = \lambda_1 + \lambda_2$	$q = \lambda_1\lambda_2$	$\Delta = (\lambda_1 - \lambda_2)^2$	Comments on λ_1, λ_2
(a) Node		$q > 0$	$\Delta \geqq 0$	Real, same sign
(b) Saddle point		$q < 0$		Real, opposite signs
(c) Center	$p = 0$	$q > 0$		Pure imaginary
(d) Spiral point	$p \neq 0$		$\Delta < 0$	Complex, not pure imaginary

Stability

Critical points may also be classified in terms of their stability. Stability concepts are basic in engineering and other applications. They are suggested by physics, where **stability** means, roughly speaking, that a small change (a small disturbance) of a physical system at some instant changes the behavior of the system only slightly at all future times t. For critical points, the following concepts are appropriate.

DEFINITIONS

Stable, Unstable, Stable and Attractive

A critical point P_0 of (1) is called **stable**[2] if, roughly, all trajectories of (1) that at some instant are close to P_0 remain close to P_0 at all future times; precisely: if for every disk D_ϵ of radius $\epsilon > 0$ with center P_0 there is a disk D_δ of radius $\delta > 0$ with center P_0 such that every trajectory of (1) that has a point P_1 (corresponding to $t = t_1$, say) in D_δ has all its points corresponding to $t \geqq t_1$ in D_ϵ. See Fig. 90.

P_0 is called **unstable** if P_0 is not stable.

P_0 is called **stable and attractive** (or *asymptotically stable*) if P_0 is stable and every trajectory that has a point in D_δ approaches P_0 as $t \to \infty$. See Fig. 91.

Classification criteria for critical points in terms of stability are given in Table 4.2. Both tables are summarized in the **stability chart** in Fig. 92. In this chart region of instability is dark blue.

[2] In the sense of the Russian mathematician ALEXANDER MICHAILOVICH LJAPUNOV (1857–1918), whose work was fundamental in stability theory for ODEs. This is perhaps the most appropriate definition of stability (and the only we shall use), but there are others, too.

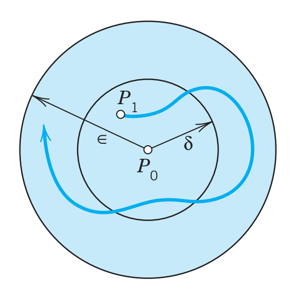


Fig. 90. Stable critical point P_0 of (1) (The trajectory initiating at P_1 stays in the disk of radius ϵ.)

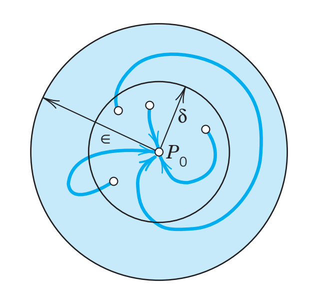


Fig. 91. Stable and attractive critical point P_0 of (1)

Table 4.2 Stability Criteria for Critical Points

Type of Stability	$p = \lambda_1 + \lambda_2$		$q = \lambda_1\lambda_2$
(a) Stable and attractive	$p < 0$		$q > 0$
(b) Stable	$p \leqq 0$		$q > 0$
(c) Unstable	$p > 0$	OR	$q < 0$

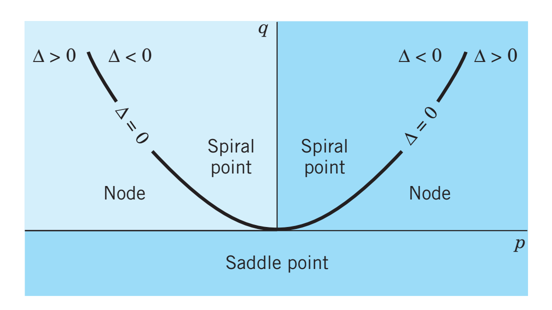


Fig. 92. Stability chart of the system (1) with p, q, Δ defined in (5). Stable and attractive: The second quadrant without the q-axis. Stability also on the positive q-axis (which corresponds to centers). Unstable: Dark blue region

We indicate how the criteria in Tables 4.1 and 4.2 are obtained. If $q = \lambda_1\lambda_2 > 0$, both of the eigenvalues are positive or both are negative or complex conjugates. If also $p = \lambda_1 + \lambda_2 < 0$, both are negative or have a negative real part. Hence P_0 is stable and attractive. The reasoning for the other two lines in Table 4.2 is similar.

If $\Delta < 0$, the eigenvalues are complex conjugates, say, $\lambda_1 = \alpha + i\beta$ and $\lambda_2 = \alpha - i\beta$. If also $p = \lambda_1 + \lambda_2 = 2\alpha < 0$, this gives a spiral point that is stable and attractive. If $p = 2\alpha > 0$, this gives an unstable spiral point.

If $p = 0$, then $\lambda_2 = -\lambda_1$ and $q = \lambda_1\lambda_2 = -\lambda_1^2$. If also $q > 0$, then $\lambda_1^2 = -q < 0$, so that λ_1, and thus λ_2, must be pure imaginary. This gives periodic solutions, their trajectories being closed curves around P_0, which is a center.

EXAMPLE 1 Application of the Criteria in Tables 4.1 and 4.2

In Example 1, Sec 4.3, we have $\mathbf{y}' = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix} \mathbf{y}$, $p = -6$, $q = 8$, $\Delta = 4$, a node by Table 4.1(a), which is stable and attractive by Table 4.2(a). ■

EXAMPLE 2 Free Motions of a Mass on a Spring

What kind of critical point does $my'' + cy' + ky = 0$ in Sec. 2.4 have?

Solution. Division by m gives $y'' = -(k/m)y - (c/m)y'$. To get a system, set $y_1 = y$, $y_2 = y'$ (see Sec. 4.1). Then $y_2' = y'' = -(k/m)y_1 - (c/m)y_2$. Hence

$$\mathbf{y}' = \begin{bmatrix} 0 & 1 \\ -k/m & -c/m \end{bmatrix} \mathbf{y}, \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -k/m & -c/m - \lambda \end{vmatrix} = \lambda^2 + \frac{c}{m}\lambda + \frac{k}{m} = 0.$$

We see that $p = -c/m$, $q = k/m$, $\Delta = (c/m)^2 - 4k/m$. From this and Tables 4.1 and 4.2 we obtain the following results. Note that in the last three cases the discriminant Δ plays an essential role.

No damping. $c = 0, p = 0, q > 0$, a center.
Underdamping. $c^2 < 4mk, p < 0, q > 0, \Delta < 0$, a stable and attractive spiral point.
Critical damping. $c^2 = 4mk, p < 0, q > 0, \Delta = 0$, a stable and attractive node.
Overdamping. $c^2 > 4mk, p < 0, q > 0, \Delta > 0$, a stable and attractive node. ■

PROBLEM SET 4.4

1–10 TYPE AND STABILITY OF CRITICAL POINT

Determine the type and stability of the critical point. Then find a real general solution and sketch or graph some of the trajectories in the phase plane. Show the details of your work.

1. $y_1' = y_1$
$y_2' = 2y_2$

2. $y_1' = -4y_1$
$y_2' = -3y_2$

3. $y_1' = y_2$
$y_2' = -9y_1$

4. $y_1' = 2y_1 + y_2$
$y_2' = 5y_1 - 2y_2$

5. $y_1' = -2y_1 + 2y_2$
$y_2' = -2y_1 - 2y_2$

6. $y_1' = -6y_1 - y_2$
$y_2' = -9y_1 - 6y_2$

7. $y_1' = y_1 + 2y_2$
$y_2' = 2y_1 + y_2$

8. $y_1' = -y_1 + 4y_2$
$y_2' = 3y_1 - 2y_2$

9. $y_1' = 4y_1 + y_2$
$y_2' = 4y_1 + 4y_2$

10. $y_1' = y_2$
$y_2' = -5y_1 - 2y_2$

11–18 TRAJECTORIES OF SYSTEMS AND SECOND-ORDER ODEs. CRITICAL POINTS

11. Damped oscillations. Solve $y'' + 2y' + 2y = 0$. What kind of curves are the trajectories?

12. Harmonic oscillations. Solve $y'' + \frac{1}{9}y = 0$. Find the trajectories. Sketch or graph some of them.

13. Types of critical points. Discuss the critical points in (10)–(13) of Sec. 4.3 by using Tables 4.1 and 4.2.

14. Transformation of parameter. What happens to the critical point in Example 1 if you introduce $\tau = -t$ as a new independent variable?

15. Perturbation of center. What happens in Example 4 of Sec. 4.3 if you change $\mathbf{A}$ to $\mathbf{A} + 0.1\mathbf{I}$, where $\mathbf{I}$ is the unit matrix?

16. Perturbation of center. If a system has a center as its critical point, what happens if you replace the matrix $\mathbf{A}$ by $\tilde{\mathbf{A}} = \mathbf{A} + k\mathbf{I}$ with any real number $k \neq 0$ (representing measurement errors in the diagonal entries)?

17. Perturbation. The system in Example 4 in Sec. 4.3 has a center as its critical point. Replace each a_{jk} in Example 4, Sec. 4.3, by $a_{jk} + b$. Find values of b such that you get **(a)** a saddle point, **(b)** a stable and attractive node, **(c)** a stable and attractive spiral, **(d)** an unstable spiral, **(e)** an unstable node.

18. CAS EXPERIMENT. Phase Portraits. Graph phase portraits for the systems in Prob. 17 with the values of b suggested in the answer. Try to illustrate how the phase portrait changes "continuously" under a continuous change of b.

19. WRITING PROBLEM. Stability. Stability concepts are basic in physics and engineering. Write a two-part report of 3 pages each (A) on general applications in which stability plays a role (be as precise as you can), and (B) on material related to stability in this section. Use your own formulations and examples; do not copy.

20. Stability chart. Locate the critical points of the systems (10)–(14) in Sec. 4.3 and of Probs. 1, 3, 5 in this problem set on the stability chart.

4.5 Qualitative Methods for Nonlinear Systems

Qualitative methods are methods of obtaining qualitative information on solutions without actually solving a system. These methods are particularly valuable for systems whose solution by analytic methods is difficult or impossible. This is the case for many practically important **nonlinear systems**

$$\mathbf{y}' = \mathbf{f}(\mathbf{y}), \qquad \text{thus} \qquad \begin{aligned} y_1' &= f_1(y_1, y_2) \\ y_2' &= f_2(y_1, y_2). \end{aligned} \tag{1}$$

In this section we extend **phase plane methods**, as just discussed, from linear systems to nonlinear systems (1). We assume that (1) is **autonomous**, that is, the independent variable t does not occur explicitly. (All examples in the last section are autonomous.) We shall again exhibit entire families of solutions. This is an advantage over numeric methods, which give only one (approximate) solution at a time.

Concepts needed from the last section are the **phase plane** (the $y_1 y_2$-plane), **trajectories** (solution curves of (1) in the phase plane), the **phase portrait** of (1) (the totality of these trajectories), and **critical points** of (1) (points (y_1, y_2) at which both $f_1(y_1, y_2)$ and $f_2(y_1, y_2)$ are zero).

Now (1) may have several critical points. Our approach shall be to discuss one critical point after another. If a critical point P_0 is not at the origin, then, for technical convenience, we shall move this point to the origin before analyzing the point. More formally, if P_0: (a, b) is a critical point with (a, b) not at the origin $(0, 0)$, then we apply the translation

$$\tilde{y}_1 = y_1 - a, \qquad \tilde{y}_2 = y_2 - b$$

which moves P_0 to $(0, 0)$ as desired. Thus we can assume P_0 to be the origin $(0, 0)$, and for simplicity we continue to write y_1, y_2 (instead of $\tilde{y}_1, \tilde{y}_2$). We also assume that P_0 is **isolated**, that is, it is the only critical point of (1) within a (sufficiently small) disk with center at the origin. If (1) has only finitely many critical points, that is automatically true. (Explain!)

Linearization of Nonlinear Systems

How can we determine the kind and stability property of a critical point P_0: $(0, 0)$ of (1)? In most cases this can be done by **linearization** of (1) near P_0, writing (1) as $\mathbf{y}' = \mathbf{f}(\mathbf{y}) = \mathbf{A}\mathbf{y} + \mathbf{h}(\mathbf{y})$ and dropping $\mathbf{h}(\mathbf{y})$, as follows.

Since P_0 is critical, $f_1(0, 0) = 0$, $f_2(0, 0) = 0$, so that f_1 and f_2 have no constant terms and we can write

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{h}(\mathbf{y}), \qquad \text{thus} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + h_1(y_1, y_2) \\ y_2' &= a_{21}y_1 + a_{22}y_2 + h_2(y_1, y_2). \end{aligned} \tag{2}$$

A is constant (independent of t) since (1) is autonomous. One can prove the following (proof in Ref. [A7], pp. 375–388, listed in App. 1).

THEOREM 1

Linearization

If f_1 *and* f_2 *in* (1) *are continuous and have continuous partial derivatives in a neighborhood of the critical point* P_0: (0, 0), *and if* $\det \mathbf{A} \neq 0$ *in* (2), *then the kind and stability of the critical point of* (1) *are the same as those of the* **linearized system**

(3)
$$\mathbf{y}' = \mathbf{A}\mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 \\ y_2' &= a_{21}y_1 + a_{22}y_2. \end{aligned}$$

Exceptions occur if $\mathbf{A}$ *has equal or pure imaginary eigenvalues; then* (1) *may have the same kind of critical point as* (3) *or a spiral point.*

EXAMPLE 1

Free Undamped Pendulum. Linearization

Figure 93a shows a pendulum consisting of a body of mass m (the bob) and a rod of length L. Determine the locations and types of the critical points. Assume that the mass of the rod and air resistance are negligible.

Solution. ***Step 1. Setting up the mathematical model.*** Let θ denote the angular displacement, measured counterclockwise from the equilibrium position. The weight of the bob is mg (g the acceleration of gravity). It causes a restoring force $mg \sin\theta$ tangent to the curve of motion (circular arc) of the bob. By Newton's second law, at each instant this force is balanced by the force of acceleration $mL\theta''$, where $L\theta''$ is the acceleration; hence the resultant of these two forces is zero, and we obtain as the mathematical model

$$mL\theta'' + mg\sin\theta = 0.$$

Dividing this by mL, we have

(4)
$$\theta'' + k\sin\theta = 0 \qquad \left(k = \frac{g}{L}\right).$$

When θ is very small, we can approximate $\sin\theta$ rather accurately by θ and obtain as an *approximate* solution $A\cos\sqrt{k}t + B\sin\sqrt{k}t$, but the *exact* solution for any θ is not an elementary function.

Step 2. Critical points $(0, 0), (\pm 2\pi, 0), (\pm 4\pi, 0), \cdots$, ***Linearization.*** To obtain a system of ODEs, we set $\theta = y_1$, $\theta' = y_2$. Then from (4) we obtain a nonlinear system (1) of the form

(4*)
$$\begin{aligned} y_1' &= f_1(y_1, y_2) = y_2 \\ y_2' &= f_2(y_1, y_2) = -k\sin y_1. \end{aligned}$$

The right sides are both zero when $y_2 = 0$ and $\sin y_1 = 0$. This gives infinitely many critical points $(n\pi, 0)$, where $n = 0, \pm1, \pm2, \cdots$. We consider (0, 0). Since the Maclaurin series is

$$\sin y_1 = y_1 - \tfrac{1}{6}y_1^3 + - \cdots \approx y_1,$$

the linearized system at (0, 0) is

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -k & 0 \end{bmatrix}\mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= -ky_1. \end{aligned}$$

To apply our criteria in Sec. 4.4 we calculate $p = a_{11} + a_{22} = 0$, $q = \det\mathbf{A} = k = g/L\ (>0)$, and $\Delta = p^2 - 4q = -4k$. From this and Table 4.1(c) in Sec. 4.4 we conclude that (0, 0) is a center, which is always stable. Since $\sin\theta = \sin y_1$ is periodic with period 2π, the critical points $(n\pi, 0)$, $n = \pm2, \pm4, \cdots$, are all centers.

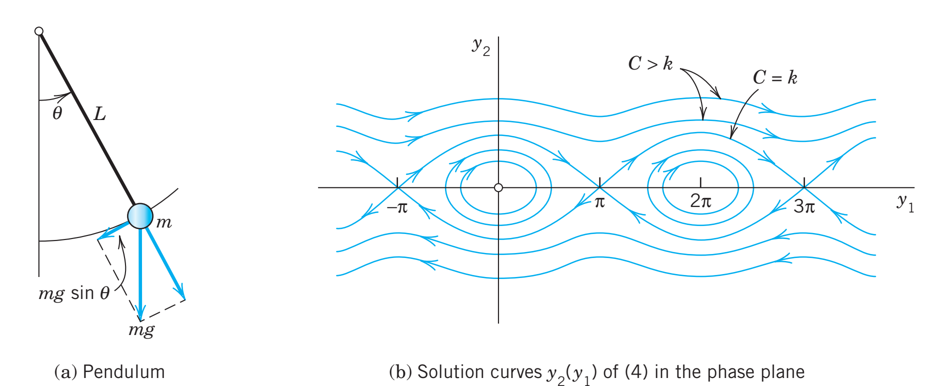
Step 3. Critical points $(\pm\pi, 0), (\pm 3\pi, 0), (\pm 5\pi, 0), \cdots$, ***Linearization.*** We now consider the critical point $(\pi, 0)$, setting $\theta - \pi = y_1$ and $(\theta - \pi)' = \theta' = y_2$. Then in (4),

$$\sin\theta = \sin(y_1 + \pi) = -\sin y_1 = -y_1 + \tfrac{1}{6}y_1^3 - + \cdots \approx -y_1$$

and the linearized system at $(\pi, 0)$ is now

$$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ k & 0 \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= k y_1. \end{aligned}$$

We see that $p = 0$, $q = -k\ (<0)$, and $\Delta = -4q = 4k$. Hence, by Table 4.1(b), this gives a saddle point, which is always unstable. Because of periodicity, the critical points $(n\pi, 0)$, $n = \pm 1, \pm 3, \cdots$, are all saddle points. These results agree with the impression we get from Fig. 93b. ■



(a) Pendulum (b) Solution curves $y_2(y_1)$ of (4) in the phase plane

Fig. 93. Example 1 (*C* will be explained in Example 4.)

EXAMPLE 2 Linearization of the Damped Pendulum Equation

To gain further experience in investigating critical points, as another practically important case, let us see how Example 1 changes when we add a damping term $c\theta'$ (damping proportional to the angular velocity) to equation (4), so that it becomes

(5) $$\theta'' + c\theta' + k \sin\theta = 0$$

where $k > 0$ and $c \geqq 0$ (which includes our previous case of no damping, $c = 0$). Setting $\theta = y_1$, $\theta' = y_2$, as before, we obtain the nonlinear system (use $\theta'' = y_2'$)

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= -k \sin y_1 - c y_2. \end{aligned}$$

We see that the critical points have the same locations as before, namely, $(0, 0)$, $(\pm\pi, 0)$, $(\pm 2\pi, 0)$, $\cdots$. We consider $(0, 0)$. Linearizing $\sin y_1 \approx y_1$ as in Example 1, we get the linearized system at $(0, 0)$

(6) $$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ -k & -c \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= -k y_1 - c y_2. \end{aligned}$$

This is identical with the system in Example 2 of Sec. 4.4, except for the (positive!) factor m (and except for the physical meaning of y_1). Hence for $c = 0$ (no damping) we have a center (see Fig. 93b), for small damping we have a spiral point (see Fig. 94), and so on.

We now consider the critical point $(\pi, 0)$. We set $\theta - \pi = y_1$, $(\theta - \pi)' = \theta' = y_2$ and linearize

$$\sin\theta = \sin(y_1 + \pi) = -\sin y_1 \approx -y_1.$$

This gives the new linearized system at $(\pi, 0)$

(6*) $$\mathbf{y}' = \mathbf{A}\mathbf{y} = \begin{bmatrix} 0 & 1 \\ k & -c \end{bmatrix} \mathbf{y}, \qquad \text{thus} \qquad \begin{aligned} y_1' &= y_2 \\ y_2' &= k y_1 - c y_2. \end{aligned}$$

For our criteria in Sec. 4.4 we calculate $p = a_{11} + a_{22} = -c$, $q = \det \mathbf{A} = -k$, and $\Delta = p^2 - 4q = c^2 + 4k$. This gives the following results for the critical point at $(\pi, 0)$.

No damping. $c = 0, p = 0, q < 0, \Delta > 0$, a saddle point. See Fig. 93b.
Damping. $c > 0, p < 0, q < 0, \Delta > 0$, a saddle point. See Fig. 94.

Since $\sin y_1$ is periodic with period 2π, the critical points $(\pm 2\pi, 0), (\pm 4\pi, 0), \cdots$ are of the same type as $(0, 0)$, and the critical points $(-\pi, 0), (\pm 3\pi, 0), \cdots$ are of the same type as $(\pi, 0)$, so that our task is finished.

Figure 94 shows the trajectories in the case of damping. What we see agrees with our physical intuition. Indeed, damping means loss of energy. Hence instead of the closed trajectories of periodic solutions in Fig. 93b we now have trajectories spiraling around one of the critical points $(0, 0), (\pm 2\pi, 0), \cdots$. Even the wavy trajectories corresponding to whirly motions eventually spiral around one of these points. Furthermore, there are no more trajectories that connect critical points (as there were in the undamped case for the saddle points). ■

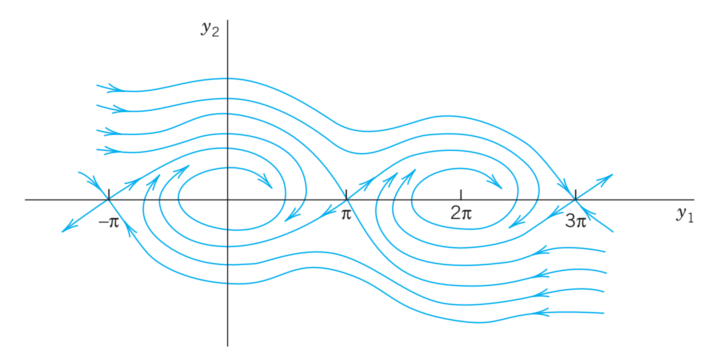


Fig. 94. Trajectories in the phase plane for the damped pendulum in Example 2

Lotka–Volterra Population Model

EXAMPLE 3 **Predator–Prey Population Model**[3]

This model concerns two species, say, rabbits and foxes, and the foxes prey on the rabbits.

Step 1. Setting up the model. We assume the following.

1. Rabbits have unlimited food supply. Hence, if there were no foxes, their number $y_1(t)$ would grow exponentially, $y_1' = ay_1$.
2. Actually, y_1 is decreased because of the kill by foxes, say, at a rate proportional to $y_1 y_2$, where $y_2(t)$ is the number of foxes. Hence $y_1' = ay_1 - by_1 y_2$, where $a > 0$ and $b > 0$.
3. If there were no rabbits, then $y_2(t)$ would exponentially decrease to zero, $y_2' = -ly_2$. However, y_2 is increased by a rate proportional to the number of encounters between predator and prey; together we have $y_2' = -ly_2 + ky_1 y_2$, where $k > 0$ and $l > 0$.

This gives the (nonlinear!) Lotka–Volterra system

$$\begin{aligned} y_1' &= f_1(y_1, y_2) = ay_1 - by_1 y_2 \\ y_2' &= f_2(y_1, y_2) = ky_1 y_2 - ly_2. \end{aligned} \tag{7}$$

[3]Introduced by ALFRED J. LOTKA (1880–1949), American biophysicist, and VITO VOLTERRA (1860–1940), Italian mathematician, the initiator of functional analysis (see [GR7] in App. 1).

Step 2. Critical point* (0, 0), *Linearization. We see from (7) that the critical points are the solutions of

$$(7^*) \qquad f_1(y_1, y_2) = y_1(a - by_2) = 0, \qquad f_2(y_1, y_2) = y_2(ky_1 - l) = 0.$$

The solutions are $(y_1, y_2) = (0, 0)$ and $\left(\frac{l}{k}, \frac{a}{b}\right)$. We consider (0, 0). Dropping $-by_1y_2$ and ky_1y_2 from (7) gives the linearized system

$$\mathbf{y}' = \begin{bmatrix} a & 0 \\ 0 & -l \end{bmatrix} \mathbf{y}.$$

Its eigenvalues are $\lambda_1 = a > 0$ and $\lambda_2 = -l < 0$. They have opposite signs, so that we get a saddle point.

Step. 3. Critical point (l/k, a/b), Linearization. We set $y_1 = \tilde{y}_1 + l/k$, $y_2 = \tilde{y}_2 + a/b$. Then the critical point $(l/k, a/b)$ corresponds to $(\tilde{y}_1, \tilde{y}_2) = (0, 0)$. Since $\tilde{y}_1' = y_1'$, $\tilde{y}_2' = y_2'$, we obtain from (7) [factorized as in (7*)]

$$\tilde{y}_1' = \left(\tilde{y}_1 + \frac{l}{k}\right)\left[a - b\left(\tilde{y}_2 + \frac{a}{b}\right)\right] = \left(\tilde{y}_1 + \frac{l}{k}\right)(-b\tilde{y}_2)$$

$$\tilde{y}_2' = \left(\tilde{y}_2 + \frac{a}{b}\right)\left[k\left(\tilde{y}_1 + \frac{l}{k}\right) - l\right] = \left(\tilde{y}_2 + \frac{a}{b}\right)k\tilde{y}_1.$$

Dropping the two nonlinear terms $-b\tilde{y}_1\tilde{y}_2$ and $k\tilde{y}_1\tilde{y}_2$, we have the linearized system

$$(7^{**}) \qquad \text{(a)} \quad \tilde{y}_1' = -\frac{lb}{k}\tilde{y}_2 \qquad\qquad \text{(b)} \quad \tilde{y}_2' = \frac{ak}{b}\tilde{y}_1.$$

The left side of (a) times the right side of (b) must equal the right side of (a) times the left side of (b),

$$\frac{ak}{b}\tilde{y}_1\tilde{y}_1' = -\frac{lb}{k}\tilde{y}_2\tilde{y}_2'. \qquad \text{By integration,} \qquad \frac{ak}{b}\tilde{y}_1^2 + \frac{lb}{k}\tilde{y}_2^2 = \text{const.}$$

This is a family of ellipses, so that the critical point $(l/k, a/b)$ of the linearized system (7**) is a center (Fig. 95). It can be shown, by a complicated analysis, that the nonlinear system (7) also has a center (rather than a spiral point) at $(l/k, a/b)$ surrounded by closed trajectories (not ellipses).

We see that the predators and prey have a cyclic variation about the critical point. Let us move counterclockwise around the ellipse, beginning at the right vertex, where the rabbits have a maximum number. Foxes are sharply increasing in number until they reach a maximum at the upper vertex, and the number of rabbits is then sharply decreasing until it reaches a minimum at the left vertex, and so on. Cyclic variations of this kind have been observed in nature, for example, for lynx and snowshoe hare near the Hudson Bay, with a cycle of about 10 years.

For models of more complicated situations and a systematic discussion, see C. W. Clark, *Mathematical Bioeconomics: The Mathematics of Conservation,* 3rd ed. Hoboken, NJ, Wiley, 2010. ■

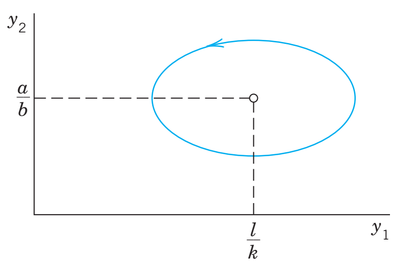


Fig. 95. Ecological equilibrium point and trajectory of the linearized Lotka–Volterra system (7**)

Transformation to a First-Order Equation in the Phase Plane

Another phase plane method is based on the idea of transforming a second-order **autonomous ODE** (an ODE in which t does not occur explicitly)

$$F(y, y', y'') = 0$$

to first order by taking $y = y_1$ as the independent variable, setting $y' = y_2$ and transforming y'' by the chain rule,

$$y'' = y_2' = \frac{dy_2}{dt} = \frac{dy_2}{dy_1}\frac{dy_1}{dt} = \frac{dy_2}{dy_1} y_2.$$

Then the ODE becomes of first order,

$$F\left(y_1, y_2, \frac{dy_2}{dy_1} y_2\right) = 0 \tag{8}$$

and can sometimes be solved or treated by direction fields. We illustrate this for the equation in Example 1 and shall gain much more insight into the behavior of solutions.

EXAMPLE 4 An ODE (8) for the Free Undamped Pendulum

If in (4) $\theta'' + k \sin\theta = 0$ we set $\theta = y_1, \theta' = y_2$ (the angular velocity) and use

$$\theta'' = \frac{dy_2}{dt} = \frac{dy_2}{dy_1}\frac{dy_1}{dt} = \frac{dy_2}{dy_1} y_2, \qquad \text{we get} \qquad \frac{dy_2}{dy_1} y_2 = -k \sin y_1.$$

Separation of variables gives $y_2\, dy_2 = -k \sin y_1\, dy_1$. By integration,

$$\tfrac{1}{2} y_2^2 = k \cos y_1 + C \qquad (C \text{ constant}). \tag{9}$$

Multiplying this by mL^2, we get

$$\tfrac{1}{2} m(Ly_2)^2 - mL^2 k \cos y_1 = mL^2 C.$$

We see that these three terms are **energies**. Indeed, y_2 is the angular velocity, so that Ly_2 is the velocity and the first term is the kinetic energy. The second term (including the minus sign) is the potential energy of the pendulum, and mL^2C is its total energy, which is constant, as expected from the law of conservation of energy, because there is no damping (no loss of energy). The type of motion depends on the total energy, hence on C, as follows.

Figure 93b shows trajectories for various values of C. These graphs continue periodically with period 2π to the left and to the right. We see that some of them are ellipse-like and closed, others are wavy, and there are two trajectories (passing through the saddle points $(n\pi, 0), n = \pm 1, \pm 3, \cdots$) that separate those two types of trajectories. From (9) we see that the smallest possible C is $C = -k$; then $y_2 = 0$, and $\cos y_1 = 1$, so that the pendulum is at rest. The pendulum will change its direction of motion if there are points at which $y_2 = \theta' = 0$. Then $k \cos y_1 + C = 0$ by (9). If $y_1 = \pi$, then $\cos y_1 = -1$ and $C = k$. Hence if $-k < C < k$, then the pendulum reverses its direction for a $|y_1| = |\theta| < \pi$, and for these values of C with $|C| < k$ the pendulum oscillates. This corresponds to the closed trajectories in the figure. However, if $C > k$, then $y_2 = 0$ is impossible and the pendulum makes a whirly motion that appears as a wavy trajectory in the $y_1 y_2$-plane. Finally, the value $C = k$ corresponds to the two "separating trajectories" in Fig. 93b connecting the saddle points. ■

The phase plane method of deriving a single first-order equation (8) may be of practical interest not only when (8) can be solved (as in Example 4) but also when a solution

is not possible and we have to utilize fields (Sec. 1.2). We illustrate this with a very famous example:

EXAMPLE 5 Self-Sustained Oscillations. Van der Pol Equation

There are physical systems such that for small oscillations, energy is fed into the system, whereas for large oscillations, energy is taken from the system. In other words, large oscillations will be damped, whereas for small oscillations there is "negative damping" (feeding of energy into the system). For physical reasons we expect such a system to approach a periodic behavior, which will thus appear as a closed trajectory in the phase plane, called a **limit cycle**. A differential equation describing such vibrations is the famous **van der Pol equation**[4]

$$y'' - \mu(1 - y^2)y' + y = 0 \qquad (\mu > 0, \text{ constant}). \tag{10}$$

It first occurred in the study of electrical circuits containing vacuum tubes. For $\mu = 0$ this equation becomes $y'' + y = 0$ and we obtain harmonic oscillations. Let $\mu > 0$. The damping term has the factor $-\mu(1 - y^2)$. This is negative for small oscillations, when $y^2 < 1$, so that we have "negative damping," is zero for $y^2 = 1$ (no damping), and is positive if $y^2 > 1$ (positive damping, loss of energy). If μ is small, we expect a limit cycle that is almost a circle because then our equation differs but little from $y'' + y = 0$. If μ is large, the limit cycle will probably look different.

Setting $y = y_1$, $y' = y_2$ and using $y'' = (dy_2/dy_1)y_2$ as in (8), we have from (10)

$$\frac{dy_2}{dy_1} y_2 - \mu(1 - y_1^2)y_2 + y_1 = 0. \tag{11}$$

The isoclines in the y_1y_2-plane (the phase plane) are the curves $dy_2/dy_1 = K = \text{const}$, that is,

$$\frac{dy_2}{dy_1} = \mu(1 - y_1^2) - \frac{y_1}{y_2} = K.$$

Solving algebraically for y_2, we see that the isoclines are given by

$$y_2 = \frac{y_1}{\mu(1 - y_1^2) - K} \qquad \text{(Figs. 96, 97).}$$

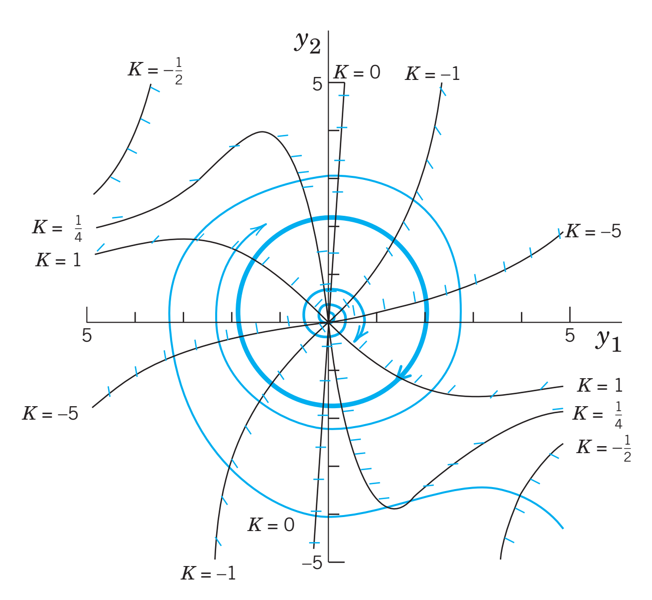


Fig. 96. Direction field for the van der Pol equation with $\mu = 0.1$ in the phase plane, showing also the limit cycle and two trajectories. See also Fig. 8 in Sec. 1.2

[4]BALTHASAR VAN DER POL (1889–1959), Dutch physicist and engineer.

Figure 96 shows some isoclines when μ is small, $\mu = 0.1$, the limit cycle (almost a circle), and two (blue) trajectories approaching it, one from the outside and the other from the inside, of which only the initial portion, a small spiral, is shown. Due to this approach by trajectories, a limit cycle differs conceptually from a closed curve (a trajectory) surrounding a center, which is not approached by trajectories. For larger μ the limit cycle no longer resembles a circle, and the trajectories approach it more rapidly than for smaller μ. Figure 97 illustrates this for $\mu = 1$. ■

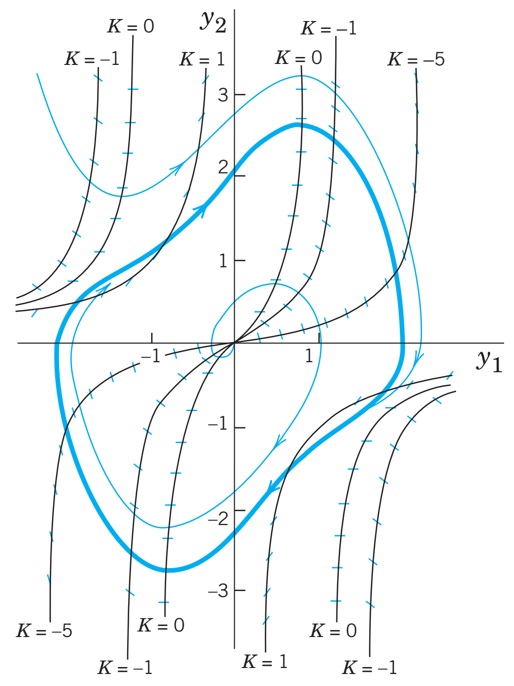


Fig. 97. Direction field for the van der Pol equation with $\mu = 1$ in the phase plane, showing also the limit cycle and two trajectories approaching it

PROBLEM SET 4.5

1. Pendulum. To what state (position, speed, direction of motion) do the four points of intersection of a closed trajectory with the axes in Fig. 93b correspond? The point of intersection of a wavy curve with the y_2-axis?

2. Limit cycle. What is the essential difference between a limit cycle and a closed trajectory surrounding a center?

3. CAS EXPERIMENT. Deformation of Limit Cycle. Convert the van der Pol equation to a system. Graph the limit cycle and some approaching trajectories for $\mu = 0.2, 0.4, 0.6, 0.8, 1.0, 1.5, 2.0$. Try to observe how the limit cycle changes its form continuously if you vary μ continuously. Describe in words how the limit cycle is deformed with growing μ.

4–8 CRITICAL POINTS. LINEARIZATION

Find the location and type of all critical points by linearization. Show the details of your work.

4. $y_1' = 4y_1 - y_1^2$, $y_2' = y_2$

5. $y_1' = y_2$, $y_2' = -y_1 + \frac{1}{2}y_1^2$

6. $y_1' = y_2$, $y_2' = -y_1 - y_1^2$

7. $y_1' = -y_1 + y_2 - y_2^2$, $y_2' = -y_1 - y_2$

8. $y_1' = y_2 - y_2^2$, $y_2' = y_1 - y_1^2$

9–13 CRITICAL POINTS OF ODEs

Find the location and type of all critical points by first converting the ODE to a system and then linearizing it.

9. $y'' - 9y + y^3 = 0$

10. $y'' + y - y^3 = 0$

11. $y'' + \cos y = 0$

12. $y'' + 9y + y^2 = 0$

13. $y'' + \sin y = 0$

14. TEAM PROJECT. Self-sustained oscillations.
(a) Van der Pol equation. Determine the type of the critical point at $(0, 0)$ when $\mu > 0, \mu = 0, \mu < 0$.
(b) Rayleigh equation. Show that the Rayleigh equation[5]

$$Y'' - \mu(1 - \tfrac{1}{3}Y'^2)Y' + Y = 0 \quad (\mu > 0)$$

also describes self-sustained oscillations and that by differentiating it and setting $y = Y'$ one obtains the van der Pol equation.
(c) Duffing equation. The Duffing equation is

$$y'' + \omega_0^2 y + \beta y^3 = 0$$

where usually $|\beta|$ is small, thus characterizing a small deviation of the restoring force from linearity. $\beta > 0$ and $\beta < 0$ are called the cases of a *hard spring* and a *soft spring*, respectively. Find the equation of the trajectories in the phase plane. (Note that for $\beta > 0$ all these curves are closed.)

15. Trajectories. Write the ODE $y'' - 4y + y^3 = 0$ as a system, solve it for y_2 as a function of y_1, and sketch or graph some of the trajectories in the phase plane.

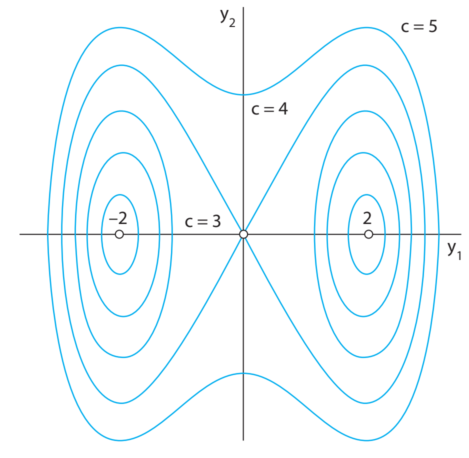


Fig. 98. Trajectories in Problem 15

4.6 Nonhomogeneous Linear Systems of ODEs

In this section, the last one of Chap. 4, we discuss methods for solving nonhomogeneous linear systems of ODEs

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} \tag{1}$$

(see Sec. 4.2)

where the vector $\mathbf{g}(t)$ is not identically zero. We assume $\mathbf{g}(t)$ and the entries of the $n \times n$ matrix $\mathbf{A}(t)$ to be continuous on some interval J of the t-axis. From a general solution $\mathbf{y}^{(h)}(t)$ of the homogeneous system $\mathbf{y}' = \mathbf{A}\mathbf{y}$ on J and a **particular solution** $\mathbf{y}^{(p)}(t)$ of (1) on J [i.e., a solution of (1) containing no arbitrary constants], we get a solution of (1),

$$\mathbf{y} = \mathbf{y}^{(h)} + \mathbf{y}^{(p)}. \tag{2}$$

$\mathbf{y}$ is called a **general solution** of (1) on J because it includes every solution of (1) on J. This follows from Theorem 2 in Sec. 4.2 (see Prob. 1 of this section).

Having studied homogeneous linear systems in Secs. 4.1–4.4, our present task will be to explain methods for obtaining particular solutions of (1). We discuss the method of

[5] LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), English physicist and mathematician, professor at Cambridge and London, known by his important contributions to the theory of waves, elasticity theory, hydrodynamics, and various other branches of applied mathematics and theoretical physics. In 1904 he was awarded the Nobel Prize in physics.

undetermined coefficients and the method of the variation of parameters; these have counterparts for a single ODE, as we know from Secs. 2.7 and 2.10.

Method of Undetermined Coefficients

Just as for a single ODE, this method is suitable if the entries of $\mathbf{A}$ are constants and the components of $\mathbf{g}$ are constants, positive integer powers of t, exponential functions, or cosines and sines. In such a case a particular solution $\mathbf{y}^{(p)}$ is assumed in a form similar to $\mathbf{g}$; for instance, $\mathbf{y}^{(p)} = \mathbf{u} + \mathbf{v}t + \mathbf{w}t^2$ if $\mathbf{g}$ has components quadratic in t, with $\mathbf{u}$, $\mathbf{v}$, $\mathbf{w}$ to be determined by substitution into (1). This is similar to Sec. 2.7, except for the Modification Rule. It suffices to show this by an example.

EXAMPLE 1 Method of Undetermined Coefficients. Modification Rule

Find a general solution of

$$\text{(3)} \qquad \mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} = \begin{bmatrix} -3 & 1 \\ 1 & -3 \end{bmatrix}\mathbf{y} + \begin{bmatrix} -6 \\ 2 \end{bmatrix} e^{-2t}.$$

Solution. A general equation of the homogeneous system is (see Example 1 in Sec. 4.3)

$$\text{(4)} \qquad \mathbf{y}^{(h)} = c_1\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t}.$$

Since $\lambda = -2$ is an eigenvalue of $\mathbf{A}$, the function e^{-2t} on the right side also appears in $\mathbf{y}^{(h)}$, and we must apply the Modification Rule by setting

$$\mathbf{y}^{(p)} = \mathbf{u}te^{-2t} + \mathbf{v}e^{-2t} \qquad \text{(rather than } \mathbf{u}e^{-2t}\text{)}.$$

Note that the first of these two terms is the analog of the modification in Sec. 2.7, but it would not be sufficient here. (Try it.) By substitution,

$$\mathbf{y}^{(p)\prime} = \mathbf{u}e^{-2t} - 2\mathbf{u}te^{-2t} - 2\mathbf{v}e^{-2t} = \mathbf{A}\mathbf{u}te^{-2t} + \mathbf{A}\mathbf{v}e^{-2t} + \mathbf{g}.$$

Equating the te^{-2t}-terms on both sides, we have $-2\mathbf{u} = \mathbf{A}\mathbf{u}$. Hence $\mathbf{u}$ is an eigenvector of $\mathbf{A}$ corresponding to $\lambda = -2$; thus [see (5)] $\mathbf{u} = a[1 \quad 1]^{\mathsf{T}}$ with any $a \neq 0$. Equating the other terms gives

$$\mathbf{u} - 2\mathbf{v} = \mathbf{A}\mathbf{v} + \begin{bmatrix} -6 \\ 2 \end{bmatrix} \qquad \text{thus} \qquad \begin{bmatrix} a \\ a \end{bmatrix} - \begin{bmatrix} 2v_1 \\ 2v_2 \end{bmatrix} = \begin{bmatrix} -3v_1 + v_2 \\ v_1 - 3v_2 \end{bmatrix} + \begin{bmatrix} -6 \\ 2 \end{bmatrix}.$$

Collecting terms and reshuffling gives

$$\begin{aligned} v_1 - v_2 &= -a - 6 \\ -v_1 + v_2 &= -a + 2. \end{aligned}$$

By addition, $0 = -2a - 4$, $a = -2$, and then $v_2 = v_1 + 4$, say, $v_1 = k$, $v_2 = k + 4$, thus, $\mathbf{v} = [k \quad k+4]^{\mathsf{T}}$. We can simply choose $k = 0$. This gives the *answer*

$$\text{(5)} \qquad \mathbf{y} = \mathbf{y}^{(h)} + \mathbf{y}^{(p)} = c_1\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2\begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} 0 \\ 4 \end{bmatrix} e^{-2t}.$$

For other k we get other $\mathbf{v}$; for instance, $k = -2$ gives $\mathbf{v} = [-2 \quad 2]^{\mathsf{T}}$, so that the *answer* becomes

$$\text{(5*)} \qquad \mathbf{y} = c_1\begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2\begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2\begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{-2t}, \qquad \text{etc.} \qquad \blacksquare$$

Method of Variation of Parameters

This method can be applied to nonhomogeneous linear systems

$$\mathbf{y}' = \mathbf{A}(t)\mathbf{y} + \mathbf{g}(t) \tag{6}$$

with variable $\mathbf{A} = \mathbf{A}(t)$ and general $\mathbf{g}(t)$. It yields a particular solution $\mathbf{y}^{(p)}$ of (6) on some open interval J on the t-axis if a general solution of the homogeneous system $\mathbf{y}' = \mathbf{A}(t)\mathbf{y}$ on J is known. We explain the method in terms of the previous example.

EXAMPLE 2 Solution by the Method of Variation of Parameters

Solve (3) in Example 1.

Solution. A basis of solutions of the homogeneous system is $[e^{-2t} \quad e^{-2t}]^{\mathsf{T}}$ and $[e^{-4t} \quad -e^{-4t}]^{\mathsf{T}}$. Hence the general solution (4) of the homogeneous system may be written

$$\mathbf{y}^{(h)} = \begin{bmatrix} e^{-2t} & e^{-4t} \\ e^{-2t} & -e^{-4t} \end{bmatrix} \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \mathbf{Y}(t)\mathbf{c}. \tag{7}$$

Here, $\mathbf{Y}(t) = [\mathbf{y}^{(1)} \quad \mathbf{y}^{(2)}]^{\mathsf{T}}$ is the fundamental matrix (see Sec. 4.2). As in Sec. 2.10 we replace the constant vector $\mathbf{c}$ by a variable vector $\mathbf{u}(t)$ to obtain a particular solution

$$\mathbf{y}^{(p)} = \mathbf{Y}(t)\mathbf{u}(t).$$

Substitution into (3) $\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}$ gives

$$\mathbf{Y}'\mathbf{u} + \mathbf{Y}\mathbf{u}' = \mathbf{A}\mathbf{Y}\mathbf{u} + \mathbf{g}. \tag{8}$$

Now since $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are solutions of the homogeneous system, we have

$$\mathbf{y}^{(1)\prime} = \mathbf{A}\mathbf{y}^{(1)}, \qquad \mathbf{y}^{(2)\prime} = \mathbf{A}\mathbf{y}^{(2)}, \qquad \text{thus} \qquad \mathbf{Y}' = \mathbf{A}\mathbf{Y}.$$

Hence $\mathbf{Y}'\mathbf{u} = \mathbf{A}\mathbf{Y}\mathbf{u}$, so that (8) reduces to

$$\mathbf{Y}\mathbf{u}' = \mathbf{g}. \qquad \text{The solution is} \qquad \mathbf{u}' = \mathbf{Y}^{-1}\mathbf{g};$$

here we use that the inverse $\mathbf{Y}^{-1}$ of $\mathbf{Y}$ (Sec. 4.0) exists because the determinant of $\mathbf{Y}$ is the Wronskian W, which is not zero for a basis. Equation (9) in Sec. 4.0 gives the form of $\mathbf{Y}^{-1}$,

$$\mathbf{Y}^{-1} = \frac{1}{-2e^{-6t}} \begin{bmatrix} -e^{-4t} & -e^{-4t} \\ -e^{-2t} & e^{-2t} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e^{2t} & e^{2t} \\ e^{4t} & -e^{4t} \end{bmatrix}.$$

We multiply this by $\mathbf{g}$, obtaining

$$\mathbf{u}' = \mathbf{Y}^{-1}\mathbf{g} = \frac{1}{2} \begin{bmatrix} e^{2t} & e^{2t} \\ e^{4t} & -e^{4t} \end{bmatrix} \begin{bmatrix} -6e^{-2t} \\ 2e^{-2t} \end{bmatrix} = \frac{1}{2} \begin{bmatrix} -4 \\ -8e^{2t} \end{bmatrix} = \begin{bmatrix} -2 \\ -4e^{2t} \end{bmatrix}.$$

Integration is done componentwise (just as differentiation) and gives

$$\mathbf{u}(t) = \int_0^t \begin{bmatrix} -2 \\ -4e^{2\tilde{t}} \end{bmatrix} d\tilde{t} = \begin{bmatrix} -2t \\ -2e^{2t} + 2 \end{bmatrix}$$

(where $+\,2$ comes from the lower limit of integration). From this and $\mathbf{Y}$ in (7) we obtain

$$\mathbf{Y}\mathbf{u} = \begin{bmatrix} e^{-2t} & e^{-4t} \\ e^{-2t} & -e^{-4t} \end{bmatrix} \begin{bmatrix} -2t \\ -2e^{2t} + 2 \end{bmatrix} = \begin{bmatrix} -2te^{-2t} - 2e^{-2t} + 2e^{-4t} \\ -2te^{-2t} + 2e^{-2t} - 2e^{-4t} \end{bmatrix} = \begin{bmatrix} -2t - 2 \\ -2t + 2 \end{bmatrix} e^{-2t} + \begin{bmatrix} 2 \\ -2 \end{bmatrix} e^{-4t}.$$

The last term on the right is a solution of the homogeneous system. Hence we can absorb it into $\mathbf{y}^{(h)}$. We thus obtain as a general solution of the system (3), in agreement with (5*).

$$\mathbf{y} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-2t} + c_2 \begin{bmatrix} 1 \\ -1 \end{bmatrix} e^{-4t} - 2 \begin{bmatrix} 1 \\ 1 \end{bmatrix} te^{-2t} + \begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{-2t}. \tag{9}$$

PROBLEM SET 4.6

1. Prove that (2) includes every solution of (1).

2–7 GENERAL SOLUTION

Find a general solution. Show the details of your work.

2. $y_1' = y_1 + y_2 + 10 \cos t$
$y_2' = 3y_1 - y_2 - 10 \sin t$

3. $y_1' = y_2 + e^{3t}$
$y_2' = y_1 - 3e^{3t}$

4. $y_1' = 4y_1 - 8y_2 + 2 \cosh t$
$y_2' = 2y_1 - 6y_2 + \cosh t + 2 \sinh t$

5. $y_1' = 4y_1 + y_2 + 0.6t$
$y_2' = 2y_1 + 3y_2 - 2.5t$

6. $y_1' = 4y_2$
$y_2' = 4y_1 - 16t^2 + 2$

7. $y_1' = -3y_1 - 4y_2 + 11t + 15$
$y_2' = 5y_1 + 6y_2 + 3e^{-t} - 15t - 20$

8. CAS EXPERIMENT. Undetermined Coefficients. Find out experimentally how general you must choose $\mathbf{y}^{(p)}$, in particular when the components of $\mathbf{g}$ have a different form (e.g., as in Prob. 7). Write a short report, covering also the situation in the case of the modification rule.

9. Undetermined Coefficients. Explain why, in Example 1 of the text, we have some freedom in choosing the vector $\mathbf{v}$.

10–15 INITIAL VALUE PROBLEM

Solve, showing details:

10. $y_1' = -3y_1 - 4y_2 + 5e^t$
$y_2' = 5y_1 + 6y_2 - 6e^t$
$y_1(0) = 19, \quad y_2(0) = -23$

11. $y_1' = y_2 + 6e^{2t}$
$y_2' = y_1 - e^{2t}$
$y_1(0) = 1, \quad y_2(0) = 0$

12. $y_1' = y_1 + 4y_2 - t^2 + 6t$
$y_2' = y_1 + y_2 - t^2 + t - 1$
$y_1(0) = 2, \quad y_2(0) = -1$

13. $y_1' = y_2 - 5 \sin t$
$y_2' = -4y_1 + 17 \cos t$
$y_1(0) = 5, \quad y_2(0) = 2$

14. $y_1' = 4y_2 + 5e^t$
$y_2' = -y_1 - 20e^{-t}$
$y_1(0) = 1, \quad y_2(0) = 0$

15. $y_1' = y_1 + 2y_2 + e^{2t} - 2t$
$y_2' = -y_2 + 1 + t$
$y_1(0) = 1, \quad y_2(0) = -4$

16. WRITING PROJECT. Undetermined Coefficients. Write a short report in which you compare the application of the method of undetermined coefficients to a single ODE and to a system of ODEs, using ODEs and systems of your choice.

17–20 NETWORK

Find the currents in Fig. 99 (Probs. 17–19) and Fig. 100 (Prob. 20) for the following data, showing the details of your work.

17. $R_1 = 2\ \Omega, R_2 = 8\ \Omega, L = 1$ H, $C = 0.5$ F, $E = 200$ V

18. Solve Prob. 17 with $E = 440 \sin t$ V and the other data as before.

19. In Prob. 17 find the particular solution when currents and charge at $t = 0$ are zero.

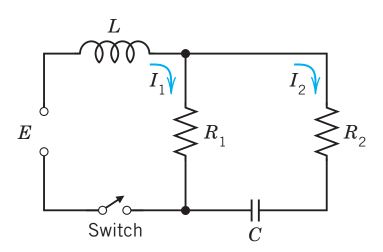


Fig. 99. Problems 17–19

20. $R_1 = 1\ \Omega$, $R_2 = 1.4\ \Omega$, $L_1 = 0.8$ H, $L_2 = 1$ H, $E = 100$ V, $I_1(0) = I_2(0) = 0$

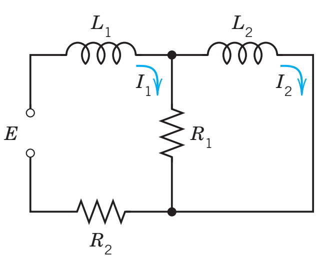


Fig. 100. Problem 20

CHAPTER 4 REVIEW QUESTIONS AND PROBLEMS

1. State some applications that can be modeled by systems of ODEs.

2. What is population dynamics? Give examples.

3. How can you transform an ODE into a system of ODEs?

4. What are qualitative methods for systems? Why are they important?

5. What is the phase plane? The phase plane method? A trajectory? The phase portrait of a system of ODEs?

6. What are critical points of a system of ODEs? How did we classify them? Why are they important?

7. What are eigenvalues? What role did they play in this chapter?

8. What does stability mean in general? In connection with critical points? Why is stability important in engineering?

9. What does linearization of a system mean?

10. Review the pendulum equations and their linearizations.

11–17 GENERAL SOLUTION. CRITICAL POINTS

Find a general solution. Determine the kind and stability of the critical point.

11. $y_1' = 2y_2$
$y_2' = 8y_1$

12. $y_1' = 5y_1$
$y_2' = y_2$

13. $y_1' = -2y_1 + 5y_2$
$y_2' = -y_1 - 6y_2$

14. $y_1' = 3y_1 + 4y_2$
$y_2' = 3y_1 + 2y_2$

15. $y_1' = -3y_1 - 2y_2$
$y_2' = -2y_1 - 3y_2$

16. $y_1' = 4y_2$
$y_2' = -4y_1$

17. $y_1' = -y_1 + 2y_2$
$y_2' = -2y_1 - y_2$

18–19 CRITICAL POINT

What kind of critical point does $\mathbf{y}' = \mathbf{A}\mathbf{y}$ have if $\mathbf{A}$ has the eigenvalues

18. -4 and 2

19. $2 + 3i,\ 2 - 3i$

20–23 NONHOMOGENEOUS SYSTEMS

Find a general solution. Show the details of your work.

20. $y_1' = 2y_1 + 2y_2 + e^t$
$y_2' = -2y_1 - 3y_2 + e^t$

21. $y_1' = 4y_2$
$y_2' = 4y_1 + 32t^2$

22. $y_1' = y_1 + y_2 + \sin t$
$y_2' = 4y_1 + y_2$

23. $y_1' = y_1 + 4y_2 - 2\cos t$
$y_2' = y_1 + y_2 - \cos t + \sin t$

24. Mixing problem. Tank T_1 in Fig. 101 initially contains 200 gal of water in which 160 lb of salt are dissolved. Tank T_2 initially contains 100 gal of pure water. Liquid is pumped through the system as indicated, and the mixtures are kept uniform by stirring. Find the amounts of salt $y_1(t)$ and $y_2(t)$ in T_1 and T_2, respectively.

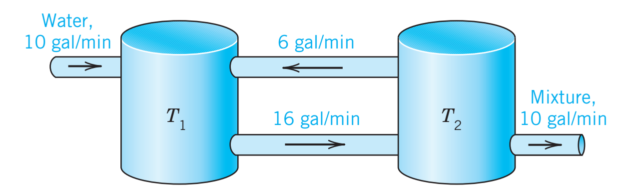


Fig. 101. Tanks in Problem 24

25. Network. Find the currents in Fig. 102 when $R = 2.5\ \Omega$, $L = 1$ H, $C = 0.04$ F, $E(t) = 169 \sin t$ V, $I_1(0) = 0$, $I_2(0) = 0$.

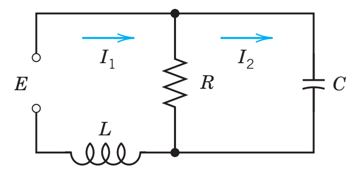


Fig. 102. Network in Problem 25

26. Network. Find the currents in Fig. 103 when $R = 1\ \Omega$, $L = 1.25$ H, $C = 0.2$ F, $I_1(0) = 1$ A, $I_2(0) = 1$ A.

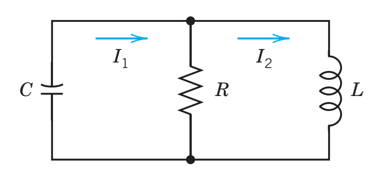


Fig. 103. Network in Problem 26

27–30 LINEARIZATION

Find the location and kind of all critical points of the given nonlinear system by linearization.

27. $y_1' = y_2$
$y_2' = y_1 - y_1^3$

28. $y_1' = \cos y_2$
$y_2' = 3y_1$

29. $y_1' = -4y_2$
$y_2' = \sin y_1$

30. $y_1' = 2y_2 + 2y_2^2$
$y_2' = -8y_1$

SUMMARY OF CHAPTER 4

Systems of ODEs. Phase Plane. Qualitative Methods

Whereas single electric circuits or single mass–spring systems are modeled by single ODEs (Chap. 2), networks of several circuits, systems of several masses and springs, and other engineering problems lead to **systems of ODEs**, involving several unknown functions $y_1(t), \cdots, y_n(t)$. Of central interest are **first-order systems** (Sec. 4.2):

$$\mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \text{in components,} \qquad \begin{aligned} y_1' &= f_1(t, y_1, \cdots, y_n) \\ &\vdots \\ y_n' &= f_n(t, y_1, \cdots, y_n), \end{aligned}$$

to which higher order ODEs and systems of ODEs can be reduced (Sec. 4.1). In this summary we let $n = 2$, so that

$$\text{(1)} \qquad \mathbf{y}' = \mathbf{f}(t, \mathbf{y}), \qquad \text{in components,} \qquad \begin{aligned} y_1' &= f_1(t, y_1, y_2) \\ y_2' &= f_2(t, y_1, y_2). \end{aligned}$$

Then we can represent solution curves as ***trajectories*** in ***the phase plane*** (the y_1y_2-plane), investigate their totality [the "***phase portrait***" of (1)], and study the kind and ***stability*** of the ***critical points*** (points at which both f_1 and f_2 are zero), and classify them as ***nodes, saddle points, centers,*** or ***spiral points*** (Secs. 4.3, 4.4). These phase plane methods are **qualitative**; with their use we can discover various general properties of solutions without actually solving the system. They are primarily used for **autonomous systems**, that is, systems in which t does not occur explicitly.

A **linear system** is of the form

$$\text{(2)} \qquad \mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g}, \quad \text{where} \quad \mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \quad \mathbf{g} = \begin{bmatrix} g_1 \\ g_2 \end{bmatrix}.$$

If $\mathbf{g} = \mathbf{0}$, the system is called **homogeneous** and is of the form

$$\text{(3)} \qquad \mathbf{y}' = \mathbf{A}\mathbf{y}.$$

If $a_{11}, \cdots, a_{22}$ are constants, it has solutions $\mathbf{y} = \mathbf{x}e^{\lambda t}$, where λ is a solution of the quadratic equation

$$\begin{vmatrix} a_{11} - \lambda & a_{12} \\ a_{21} & a_{22} - \lambda \end{vmatrix} = (a_{11} - \lambda)(a_{22} - \lambda) - a_{12}a_{21} = 0$$

and $\mathbf{x} \neq \mathbf{0}$ has components x_1, x_2 determined up to a multiplicative constant by

$$(a_{11} - \lambda)x_1 + a_{12}x_2 = 0.$$

(These λ's are called the **eigenvalues** and these vectors $\mathbf{x}$ **eigenvectors** of the matrix $\mathbf{A}$. Further explanation is given in Sec. 4.0.)

A system (2) with $\mathbf{g} \neq \mathbf{0}$ is called **nonhomogeneous**. Its general solution is of the form $\mathbf{y} = \mathbf{y}_h + \mathbf{y}_p$, where $\mathbf{y}_h$ is a general solution of (3) and $\mathbf{y}_p$ a particular solution of (2). Methods of determining the latter are discussed in Sec. 4.6.

The discussion of critical points of linear systems based on eigenvalues is summarized in Tables 4.1 and 4.2 in Sec. 4.4. It also applies to nonlinear systems if the latter are first linearized. The key theorem for this is Theorem 1 in Sec. 4.5, which also includes three famous applications, namely the pendulum and van der Pol equations and the Lotka–Volterra predator–prey population model.

Series Solutions of ODEs. Special Functions

In the previous chapters, we have seen that linear ODEs with *constant coefficients* can be solved by algebraic methods, and that their solutions are elementary functions known from calculus. For ODEs with *variable coefficients* the situation is more complicated, and their solutions may be nonelementary functions. *Legendre's, Bessel's*, and the *hypergeometric equations* are important ODEs of this kind. Since these ODEs and their solutions, the *Legendre polynomials, Bessel functions*, and *hypergeometric functions*, play an important role in engineering modeling, we shall consider the two standard methods for solving such ODEs.

The first method is called the **power series method** because it gives solutions in the form of a power series $a_0 + a_1x + a_2x^2 + a_3x^3 + \cdots$.

The second method is called the **Frobenius method** and generalizes the first; it gives solutions in power series, multiplied by a logarithmic term $\ln x$ or a fractional power x^r, in cases such as Bessel's equation, in which the first method is not general enough.

All those more advanced solutions and various other functions not appearing in calculus are known as *higher functions* or **special functions**, which has become a technical term. Each of these functions is important enough to give it a name and investigate its properties and relations to other functions in great detail (take a look into Refs. [GenRef1], [GenRef10], or [All] in App. 1). Your CAS knows practically all functions you will ever need in industry or research labs, but it is up to you to find your way through this vast terrain of formulas. The present chapter may give you some help in this task.

COMMENT. You can study this chapter directly after Chap. 2 because it needs no material from Chaps. 3 or 4.

Prerequisite: Chap. 2.
Section that may be omitted in a shorter course: 5.5.
References and Answers to Problems: App. 1 Part A, and App. 2.

5.1 Power Series Method

The **power series method** is the standard method for solving linear ODEs with ***variable*** coefficients. It gives solutions in the form of power series. These series can be used for computing values, graphing curves, proving formulas, and exploring properties of solutions, as we shall see. In this section we begin by explaining the idea of the power series method.

From calculus we remember that a **power series** (in powers of $x - x_0$) is an infinite series of the form

$$\sum_{m=0}^{\infty} a_m(x - x_0)^m = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots. \quad (1)$$

Here, x is a variable. $a_0, a_1, a_2, \cdots$ are constants, called the **coefficients** of the series. x_0 is a constant, called the **center** of the series. In particular, if $x_0 = 0$, we obtain a **power series in powers of *x***

$$\sum_{m=0}^{\infty} a_m x^m = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots. \quad (2)$$

We shall assume that all variables and constants are real.

We note that the term "power series" usually refers to a series of the form (1) [or (2)] but ***does not include*** series of negative or fractional powers of x. We use m as the summation letter, reserving n as a standard notation in the Legendre and Bessel equations for integer values of the parameter.

EXAMPLE 1 **Familiar Power Series** are the Maclaurin series

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m = 1 + x + x^2 + \cdots \qquad (|x| < 1, \textbf{ geometric series})$$

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots$$

$$\cos x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{(2m)!} = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - + \cdots$$

$$\sin x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{(2m+1)!} = x - \frac{x^3}{3!} + \frac{x^5}{5!} - + \cdots.$$

■

Idea and Technique of the Power Series Method

The idea of the power series method for solving linear ODEs seems natural, once we know that the most important ODEs in applied mathematics have solutions of this form. We explain the idea by an ODE that can readily be solved otherwise.

EXAMPLE 2 **Power Series Solution.** Solve $y' - y = 0$.

Solution. In the first step we insert

$$y = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \cdots = \sum_{m=0}^{\infty} a_m x^m \quad (2)$$

and the series obtained by termwise differentiation

$$y' = a_1 + 2a_2 x + 3a_3 x^2 + \cdots = \sum_{m=1}^{\infty} m a_m x^{m-1} \tag{3}$$

into the ODE:

$$(a_1 + 2a_2 x + 3a_3 x^2 + \cdots) - (a_0 + a_1 x + a_2 x^2 + \cdots) = 0.$$

Then we collect like powers of x, finding

$$(a_1 - a_0) + (2a_2 - a_1)x + (3a_3 - a_2)x^2 + \cdots = 0.$$

Equating the coefficient of each power of x to zero, we have

$$a_1 - a_0 = 0, \qquad 2a_2 - a_1 = 0, \qquad 3a_3 - a_2 = 0, \cdots.$$

Solving these equations, we may express $a_1, a_2, \cdots$ in terms of a_0, which remains arbitrary:

$$a_1 = a_0, \qquad a_2 = \frac{a_1}{2} = \frac{a_0}{2!}, \qquad a_3 = \frac{a_2}{3} = \frac{a_0}{3!}, \cdots.$$

With these values of the coefficients, the series solution becomes the familiar general solution

$$y = a_0 + a_0 x + \frac{a_0}{2!}x^2 + \frac{a_0}{3!}x^3 + \cdots = a_0\left(1 + x + \frac{x^2}{2!} + \frac{x^3}{3!}\right) = a_0 e^x.$$

Test your comprehension by solving $y'' + y = 0$ by power series. You should get the result $y = a_0 \cos x + a_1 \sin x$. ■

We now describe the method in general and justify it after the next example. For a given ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{4}$$

we first represent $p(x)$ and $q(x)$ by power series in powers of x (or of $x - x_0$ if solutions in powers of $x - x_0$ are wanted). Often $p(x)$ and $q(x)$ are polynomials, and then nothing needs to be done in this first step. Next we assume a solution in the form of a power series (2) with unknown coefficients and insert it as well as (3) and

$$y'' = 2a_2 + 3 \cdot 2a_3 x + 4 \cdot 3a_4 x^2 + \cdots = \sum_{m=2}^{\infty} m(m-1)a_m x^{m-2} \tag{5}$$

into the ODE. Then we collect like powers of x and equate the sum of the coefficients of each occurring power of x to zero, starting with the constant terms, then taking the terms containing x, then the terms in x^2, and so on. This gives equations from which we can determine the unknown coefficients of (3) successively.

EXAMPLE 3 **A Special Legendre Equation.** The ODE

$$(1 - x^2)y'' - 2xy' + 2y = 0$$

occurs in models exhibiting spherical symmetry. Solve it.

Solution. Substitute (2), (3), and (5) into the ODE. $(1 - x^2)y''$ gives two series, one for y'' and one for $-x^2y''$. In the term $-2xy'$ use (3) and in $2y$ use (2). Write like powers of x vertically aligned. This gives

$$\begin{aligned} y'' &= 2a_2 + 6a_3x + 12a_4x^2 + 20a_5x^3 + 30a_6x^4 + \cdots \\ -x^2y'' &= \qquad\qquad\quad - 2a_2x^2 - 6a_3x^3 - 12a_4x^4 - \cdots \\ -2xy' &= \qquad - 2a_1x - 4a_2x^2 - 6a_3x^3 - 8a_4x^4 - \cdots \\ 2y &= 2a_0 + 2a_1x + 2a_2x^2 + 2a_3x^3 + 2a_4x^4 + \cdots. \end{aligned}$$

Add terms of like powers of x. For each power $x^0, x, x^2, \cdots$ equate the sum obtained to zero. Denote these sums by [0] (constant terms), [1] (first power of x), and so on:

Sum	Power	Equations	
[0]	$[x^0]$	$a_2 = -a_0$	
[1]	$[x]$	$a_3 = 0$	
[2]	$[x^2]$	$12a_4 = 4a_2,$	$a_4 = \frac{4}{12}a_2 = -\frac{1}{3}a_0$
[3]	$[x^3]$	$a_5 = 0$	since $a_3 = 0$
[4]	$[x^4]$	$30a_6 = 18a_4,$	$a_6 = \frac{18}{30}a_4 = \frac{18}{30}(-\frac{1}{3})a_0 = -\frac{1}{5}a_0.$

This gives the solution

$$y = a_1x + a_0(1 - x^2 - \tfrac{1}{3}x^4 - \tfrac{1}{5}x^6 - \cdots).$$

a_0 and a_1 remain arbitrary. Hence, this is a general solution that consists of two solutions: x and $1 - x^2 - \frac{1}{3}x^4 - \frac{1}{5}x^6 - \cdots$. These two solutions are members of families of functions called *Legendre polynomials* $P_n(x)$ and *Legendre functions* $Q_n(x)$; here we have $x = P_1(x)$ and $1 - x^2 - \frac{1}{3}x^4 - \frac{1}{5}x^6 - \cdots = -Q_1(x)$. The minus is by convention. The index 1 is called the *order* of these two functions and here the order is 1. More on Legendre polynomials in the next section. ■

Theory of the Power Series Method

The ***n*th partial sum** of (1) is

$$s_n(x) = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots + a_n(x - x_0)^n \tag{6}$$

where $n = 0, 1, \cdots$. If we omit the terms of s_n from (1), the remaining expression is

$$R_n(x) = a_{n+1}(x - x_0)^{n+1} + a_{n+2}(x - x_0)^{n+2} + \cdots. \tag{7}$$

This expression is called the **remainder** *of* (1) *after the term* $a_n(x - x_0)^n$.

For example, in the case of the geometric series

$$1 + x + x^2 + \cdots + x^n + \cdots$$

we have

$$\begin{aligned} s_0 &= 1, & R_0 &= x + x^2 + x^3 + \cdots, \\ s_1 &= 1 + x, & R_1 &= x^2 + x^3 + x^4 + \cdots, \\ s_2 &= 1 + x + x^2, & R_2 &= x^3 + x^4 + x^5 + \cdots, \qquad \text{etc.} \end{aligned}$$

In this way we have now associated with (1) the sequence of the partial sums $s_0(x), s_1(x), s_2(x), \cdots$. If for some $x = x_1$ this sequence converges, say,

$$\lim_{n\to\infty} s_n(x_1) = s(x_1),$$

then the series (1) is called **convergent** *at* $x = x_1$, the number $s(x_1)$ is called the **value** or *sum* of (1) at x_1, and we write

$$s(x_1) = \sum_{m=0}^{\infty} a_m(x_1 - x_0)^m.$$

Then we have for every n,

$$s(x_1) = s_n(x_1) + R_n(x_1). \tag{8}$$

If that sequence diverges at $x = x_1$, the series (1) is called **divergent** at $x = x_1$.

In the case of convergence, for any positive ϵ there is an N (depending on ϵ) such that, by (8)

$$|R_n(x_1)| = |s(x_1) - s_n(x_1)| < \epsilon \qquad \text{for all } n > N. \tag{9}$$

Geometrically, this means that all $s_n(x_1)$ with $n > N$ lie between $s(x_1) - \epsilon$ and $s(x_1) + \epsilon$ (Fig. 104). Practically, this means that in the case of convergence we can approximate the sum $s(x_1)$ of (1) at x_1 by $s_n(x_1)$ as accurately as we please, by taking n large enough.

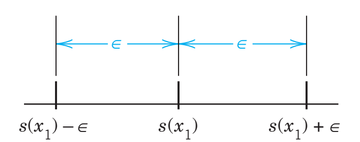


Fig. 104. Inequality (9)

Where does a power series converge? Now if we choose $x = x_0$ in (1), the series reduces to the single term a_0 because the other terms are zero. Hence the series converges at x_0. In some cases this may be the only value of x for which (1) converges. If there are other values of x for which the series converges, these values form an interval, the **convergence interval**. This interval may be finite, as in Fig. 105, with midpoint x_0. Then the series (1) converges for all x in the interior of the interval, that is, for all x for which

$$|x - x_0| < R \tag{10}$$

and diverges for $|x - x_0| > R$. The interval may also be infinite, that is, the series may converge for all x.

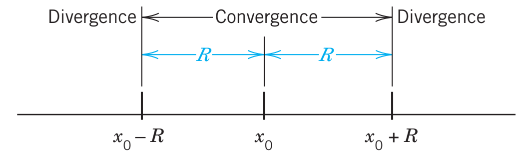


Fig. 105. Convergence interval (10) of a power series with center x_0

The quantity R in Fig. 105 is called the **radius of convergence** (because for a *complex* power series it is the radius of *disk* of convergence). If the series converges for all x, we set $R = \infty$ (and $1/R = 0$).

The radius of convergence can be determined from the coefficients of the series by means of each of the formulas

(11) $$\text{(a)}\quad R = 1\Big/\lim_{m\to\infty}\sqrt[m]{|a_m|} \qquad \text{(b)}\quad R = 1\Big/\lim_{m\to\infty}\left|\frac{a_{m+1}}{a_m}\right|$$

provided these limits exist and are not zero. [If these limits are infinite, then (1) converges only at the center x_0.]

EXAMPLE 4 Convergence Radius $R = \infty, 1, 0$

For all three series let $m \to \infty$

$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!} = 1 + x + \frac{x^2}{2!} + \cdots, \qquad \left|\frac{a_{m+1}}{a_m}\right| = \frac{1/(m+1)!}{1/m!} = \frac{1}{m+1} \to 0, \qquad R = \infty$$

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m = 1 + x + x^2 + \cdots, \qquad \left|\frac{a_{m+1}}{a_m}\right| = \frac{1}{1} = 1, \qquad R = 1$$

$$\sum_{m=0}^{\infty} m!x^m = 1 + x + 2x^2 + \cdots, \qquad \left|\frac{a_{m+1}}{a_m}\right| = \frac{(m+1)!}{m!} = m + 1 \to \infty, \qquad R = 0.$$

Convergence for all x $(R = \infty)$ is the best possible case, convergence in some finite interval the usual, and convergence only at the center $(R = 0)$ is useless. ■

When do power series solutions exist? *Answer:* if p, q, r in the ODEs

(12) $$y'' + p(x)y' + q(x)y = r(x)$$

have power series representations (Taylor series). More precisely, a function $f(x)$ is called **analytic** at a point $x = x_0$ if it can be represented by a power series in powers of $x - x_0$ with positive radius of convergence. Using this concept, we can state the following basic theorem, in which the ODE (12) is **in standard form**, that is, it begins with the y''. If your ODE begins with, say, $h(x)y''$, divide it first by $h(x)$ and then apply the theorem to the resulting new ODE.

THEOREM 1 Existence of Power Series Solutions

If p, q, and r in (12) *are analytic at $x = x_0$, then every solution of* (12) *is analytic at $x = x_0$ and can thus be represented by a power series in powers of $x - x_0$ with radius of convergence $R > 0$.*

The proof of this theorem requires advanced complex analysis and can be found in Ref. [A11] listed in App. 1.

We mention that the radius of convergence R in Theorem 1 is at least equal to the distance from the point $x = x_0$ to the point (or points) closest to x_0 at which one of the functions p, q, r, as functions of a *complex variable*, is not analytic. (Note that that point may not lie on the x-axis but somewhere in the complex plane.)

Further Theory: Operations on Power Series

In the power series method we differentiate, add, and multiply power series, and we obtain coefficient recursions (as, for instance, in Example 3) by equating the sum of the coefficients of each occurring power of x to zero. These four operations are permissible in the sense explained in what follows. Proofs can be found in Sec. 15.3.

1. Termwise Differentiation. *A power series may be differentiated term by term.* More precisely: if

$$y(x) = \sum_{m=0}^{\infty} a_m(x - x_0)^m$$

converges for $|x - x_0| < R$, where $R > 0$, then the series obtained by differentiating term by term also converges for those x and represents the derivative y' of y for those x:

$$y'(x) = \sum_{m=1}^{\infty} ma_m(x - x_0)^{m-1} \qquad (|x - x_0| < R).$$

Similarly for the second and further derivatives.

2. Termwise Addition. *Two power series may be added term by term.* More precisely: if the series

$$\sum_{m=0}^{\infty} a_m(x - x_0)^m \quad \text{and} \quad \sum_{m=0}^{\infty} b_m(x - x_0)^m \tag{13}$$

have positive radii of convergence and their sums are $f(x)$ and $g(x)$, then the series

$$\sum_{m=0}^{\infty} (a_m + b_m)(x - x_0)^m$$

converges and represents $f(x) + g(x)$ for each x that lies in the interior of the convergence interval common to each of the two given series.

3. Termwise Multiplication. *Two power series may be multiplied term by term.* More precisely: Suppose that the series (13) have positive radii of convergence and let $f(x)$ and $g(x)$ be their sums. Then the series obtained by multiplying each term of the first series by each term of the second series and collecting like powers of $x - x_0$, that is,

$$a_0b_0 + (a_0b_1 + a_1b_0)(x - x_0) + (a_0b_2 + a_1b_1 + a_2b_0)(x - x_0)^2 + \cdots$$

$$= \sum_{m=0}^{\infty} (a_0b_m + a_1b_{m-1} + \cdots + a_mb_0)(x - x_0)^m$$

converges and represents $f(x)g(x)$ for each x in the interior of the convergence interval of each of the two given series.

4. Vanishing of All Coefficients (*"Identity Theorem for Power Series."*) *If a power series has a positive radius of convergent convergence and a sum that is identically zero throughout its interval of convergence, then each coefficient of the series must be zero.*

PROBLEM SET 5.1

1. WRITING AND LITERATURE PROJECT. Power Series in Calculus. (a) Write a review (2–3 pages) on power series in calculus. Use your own formulations and examples—do not just copy from textbooks. No proofs. **(b)** Collect and arrange Maclaurin series in a systematic list that you can use for your work.

2–5 REVIEW: RADIUS OF CONVERGENCE

Determine the radius of convergence. Show the details of your work.

2. $\sum_{m=0}^{\infty} (m+1)mx^m$

3. $\sum_{m=0}^{\infty} \frac{(-1)^m}{k^m} x^{2m}$

4. $\sum_{m=0}^{\infty} \frac{x^{2m+1}}{(2m+1)!}$

5. $\sum_{m=0}^{\infty} \left(\frac{2}{3}\right)^m x^{2m}$

6–9 SERIES SOLUTIONS BY HAND

Apply the power series method. Do this by hand, not by a CAS, to get a feel for the method, e.g., why a series may terminate, or has even powers only, etc. Show the details.

6. $(1 + x)y' = y$

7. $y' = -2xy$

8. $xy' - 3y = k\ (= \text{const})$

9. $y'' + y = 0$

10–14 SERIES SOLUTIONS

Find a power series solution in powers of x. Show the details.

10. $y'' - y' + xy = 0$

11. $y'' - y' + x^2y = 0$

12. $(1 - x^2)y'' - 2xy' + 2y = 0$

13. $y'' + (1 + x^2)y = 0$

14. $y'' - 4xy' + (4x^2 - 2)y = 0$

15. Shifting summation indices is often convenient or necessary in the power series method. Shift the index so that the power under the summation sign is x^m. Check by writing the first few terms explicity.

$$\sum_{s=2}^{\infty} \frac{s(s+1)}{s^2+1} x^{s-1}, \qquad \sum_{p=1}^{\infty} \frac{p^2}{(p+1)!} x^{p+4}$$

16–19 CAS PROBLEMS. IVPs

Solve the initial value problem by a power series. Graph the partial sums of the powers up to and including x^5. Find the value of the sum s (5 digits) at x_1.

16. $y' + 4y = 1, \quad y(0) = 1.25, \quad x_1 = 0.2$

17. $y'' + 3xy' + 2y = 0, \quad y(0) = 1, \quad y'(0) = 1,$ $x = 0.5$

18. $(1 - x^2)y'' - 2xy' + 30y = 0, \quad y(0) = 0,$ $y'(0) = 1.875, \quad x_1 = 0.5$

19. $(x - 2)y' = xy, \quad y(0) = 4, \quad x_1 = 2$

20. CAS Experiment. Information from Graphs of Partial Sums. In numerics we use partial sums of power series. To get a feel for the accuracy for various x, experiment with $\sin x$. Graph partial sums of the Maclaurin series of an increasing number of terms, describing qualitatively the "breakaway points" of these graphs from the graph of $\sin x$. Consider other Maclaurin series of your choice.

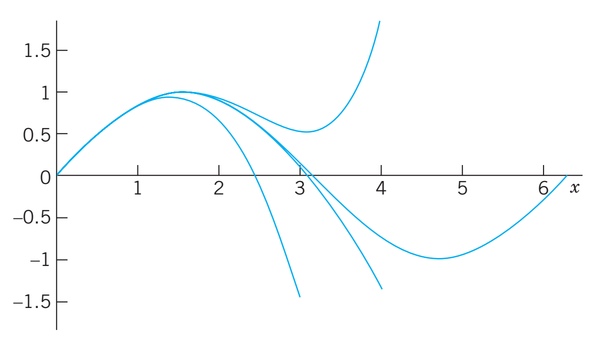


Fig. 106. CAS Experiment 20. $\sin x$ and partial sums s_3, s_5, s_7

5.2 Legendre's Equation. Legendre Polynomials $P_n(x)$

Legendre's differential equation[1]

$$(1 - x^2)y'' - 2xy' + n(n + 1)y = 0 \qquad (n \text{ constant}) \tag{1}$$

is one of the most important ODEs in physics. It arises in numerous problems, particularly in boundary value problems for spheres (take a quick look at Example 1 in Sec. 12.10).

The equation involves a **parameter** n, whose value depends on the physical or engineering problem. So (1) is actually a whole family of ODEs. For $n = 1$ we solved it in Example 3 of Sec. 5.1 (look back at it). Any solution of (1) is called a **Legendre function**. The study of these and other "higher" functions not occurring in calculus is called the **theory of special functions**. Further special functions will occur in the next sections.

Dividing (1) by $1 - x^2$, we obtain the standard form needed in Theorem 1 of Sec. 5.1 and we see that the coefficients $-2x/(1 - x^2)$ and $n(n + 1)/(1 - x^2)$ of the new equation are analytic at $x = 0$, so that we may apply the power series method. Substituting

$$y = \sum_{m=0}^{\infty} a_m x^m \tag{2}$$

and its derivatives into (1), and denoting the constant $n(n + 1)$ simply by k, we obtain

$$(1 - x^2) \sum_{m=2}^{\infty} m(m - 1)a_m x^{m-2} - 2x \sum_{m=1}^{\infty} m a_m x^{m-1} + k \sum_{m=0}^{\infty} a_m x^m = 0.$$

By writing the first expression as two separate series we have the equation

$$\sum_{m=2}^{\infty} m(m - 1)a_m x^{m-2} - \sum_{m=2}^{\infty} m(m - 1)a_m x^m - \sum_{m=1}^{\infty} 2m a_m x^m + \sum_{m=0}^{\infty} k a_m x^m = 0.$$

It may help you to write out the first few terms of each series explicitly, as in Example 3 of Sec. 5.1; or you may continue as follows. To obtain the same general power x^s in all four series, set $m - 2 = s$ (thus $m = s + 2$) in the first series and simply write s instead of m in the other three series. This gives

$$\sum_{s=0}^{\infty} (s + 2)(s + 1)a_{s+2} x^s - \sum_{s=2}^{\infty} s(s - 1)a_s x^s - \sum_{s=1}^{\infty} 2s a_s x^s + \sum_{s=0}^{\infty} k a_s x^s = 0.$$

---

[1] ADRIEN-MARIE LEGENDRE (1752–1833), French mathematician, who became a professor in Paris in 1775 and made important contributions to special functions, elliptic integrals, number theory, and the calculus of variations. His book *Éléments de géométrie* (1794) became very famous and had 12 editions in less than 30 years.

Formulas on Legendre functions may be found in Refs. [GenRef1] and [GenRef10].

(Note that in the first series the summation begins with $s = 0$.) Since this equation with the right side 0 must be an identity in x if (2) is to be a solution of (1), the sum of the coefficients of each power of x on the left must be zero. Now x^0 occurs in the first and fourth series only, and gives [remember that $k = n(n + 1)$]

$$2 \cdot 1a_2 + n(n + 1)a_0 = 0. \tag{3a}$$

x^1 occurs in the first, third, and fourth series and gives

$$3 \cdot 2a_3 + [-2 + n(n + 1)]a_1 = 0. \tag{3b}$$

The higher powers $x^2, x^3, \cdots$ occur in all four series and give

$$(s + 2)(s + 1)a_{s+2} + [-s(s - 1) - 2s + n(n + 1)]a_s = 0. \tag{3c}$$

The expression in the brackets $[\cdots]$ can be written $(n - s)(n + s + 1)$, as you may readily verify. Solving (3a) for a_2 and (3b) for a_3 as well as (3c) for a_{s+2}, we obtain the general formula

$$a_{s+2} = -\frac{(n - s)(n + s + 1)}{(s + 2)(s + 1)} a_s \qquad (s = 0, 1, \cdots). \tag{4}$$

This is called a **recurrence relation** or **recursion formula**. (Its derivation you may verify with your CAS.) It gives each coefficient in terms of the second one preceding it, except for a_0 and a_1, which are left as arbitrary constants. We find successively

$$a_2 = -\frac{n(n + 1)}{2!} a_0 \qquad\qquad a_3 = -\frac{(n - 1)(n + 2)}{3!} a_1$$

$$a_4 = -\frac{(n - 2)(n + 3)}{4 \cdot 3} a_2 \qquad\qquad a_5 = -\frac{(n - 3)(n + 4)}{5 \cdot 4} a_3$$

$$= \frac{(n - 2)n(n + 1)(n + 3)}{4!} a_0 \qquad\qquad = \frac{(n - 3)(n - 1)(n + 2)(n + 4)}{5!} a_1$$

and so on. By inserting these expressions for the coefficients into (2) we obtain

$$y(x) = a_0 y_1(x) + a_1 y_2(x) \tag{5}$$

where

$$y_1(x) = 1 - \frac{n(n + 1)}{2!} x^2 + \frac{(n - 2)n(n + 1)(n + 3)}{4!} x^4 - + \cdots \tag{6}$$

$$y_2(x) = x - \frac{(n - 1)(n + 2)}{3!} x^3 + \frac{(n - 3)(n - 1)(n + 2)(n + 4)}{5!} x^5 - + \cdots. \tag{7}$$

These series converge for $|x| < 1$ (see Prob. 4; or they may terminate, see below). Since (6) contains even powers of x only, while (7) contains odd powers of x only, the ratio y_1/y_2 is not a constant, so that y_1 and y_2 are not proportional and are thus linearly independent solutions. Hence (5) is a general solution of (1) on the interval $-1 < x < 1$.

Note that $x = \pm 1$ are the points at which $1 - x^2 = 0$, so that the coefficients of the standardized ODE are no longer analytic. So it should not surprise you that we do not get a longer convergence interval of (6) and (7), unless these series terminate after finitely many powers. In that case, the series become polynomials.

Polynomial Solutions. Legendre Polynomials $P_n(x)$

The reduction of power series to polynomials is a great advantage because then we have solutions for all x, without convergence restrictions. For special functions arising as solutions of ODEs this happens quite frequently, leading to various important families of polynomials; see Refs. [GenRef1], [GenRef10] in App. 1. For Legendre's equation this happens when the parameter n is a nonnegative integer because then the right side of (4) is zero for $s = n$, so that $a_{n+2} = 0, a_{n+4} = 0, a_{n+6} = 0, \cdots$. Hence if n is even, $y_1(x)$ reduces to a polynomial of degree n. If n is odd, the same is true for $y_2(x)$. These polynomials, multiplied by some constants, are called **Legendre polynomials** and are denoted by $P_n(x)$. The standard choice of such constants is done as follows. We choose the coefficient a_n of the highest power x^n as

$$a_n = \frac{(2n)!}{2^n(n!)^2} = \frac{1 \cdot 3 \cdot 5 \cdots (2n-1)}{n!} \qquad (n \text{ a positive integer}) \tag{8}$$

(and $a_n = 1$ if $n = 0$). Then we calculate the other coefficients from (4), solved for a_s in terms of a_{s+2}, that is,

$$a_s = -\frac{(s+2)(s+1)}{(n-s)(n+s+1)} a_{s+2} \qquad (s \leqq n-2). \tag{9}$$

The choice (8) makes $p_n(1) = 1$ for every n (see Fig. 107); this motivates (8). From (9) with $s = n - 2$ and (8) we obtain

$$a_{n-2} = -\frac{n(n-1)}{2(2n-1)} a_n = -\frac{n(n-1)}{2(2n-1)} \cdot \frac{(2n)!}{2^n(n!)^2}$$

Using $(2n)! = 2n(2n-1)(2n-2)!$ in the numerator and $n! = n(n-1)!$ and $n! = n(n-1)(n-2)!$ in the denominator, we obtain

$$a_{n-2} = -\frac{n(n-1)2n(2n-1)(2n-2)!}{2(2n-1)2^n n(n-1)!\, n(n-1)(n-2)!}.$$

$n(n-1)2n(2n-1)$ cancels, so that we get

$$a_{n-2} = -\frac{(2n-2)!}{2^n(n-1)!\,(n-2)!}.$$

Similarly,

$$a_{n-4} = -\frac{(n-2)(n-3)}{4(2n-3)} a_{n-2}$$
$$= \frac{(2n-4)!}{2^n 2!\,(n-2)!\,(n-4)!}$$

and so on, and in general, when $n - 2m \geqq 0$,

(10)
$$a_{n-2m} = (-1)^m \frac{(2n-2m)!}{2^n m!\,(n-m)!\,(n-2m)!}.$$

The resulting solution of Legendre's differential equation (1) is called the **Legendre polynomial** *of degree n* and is denoted by $P_n(x)$.

From (10) we obtain

(11)
$$P_n(x) = \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{2^n m!\,(n-m)!\,(n-2m)!} x^{n-2m}$$
$$= \frac{(2n)!}{2^n (n!)^2} x^n - \frac{(2n-2)!}{2^n 1!\,(n-1)!\,(n-2)!} x^{n-2} + - \cdots$$

where $M = n/2$ or $(n-1)/2$, whichever is an integer. The first few of these functions are (Fig. 107)

(11′)
$$P_0(x) = 1, \qquad P_1(x) = x$$
$$P_2(x) = \tfrac{1}{2}(3x^2 - 1), \qquad P_3(x) = \tfrac{1}{2}(5x^3 - 3x)$$
$$P_4(x) = \tfrac{1}{8}(35x^4 - 30x^2 + 3), \qquad P_5(x) = \tfrac{1}{8}(63x^5 - 70x^3 + 15x)$$

and so on. You may now program (11) on your CAS and calculate $P_n(x)$ as needed.

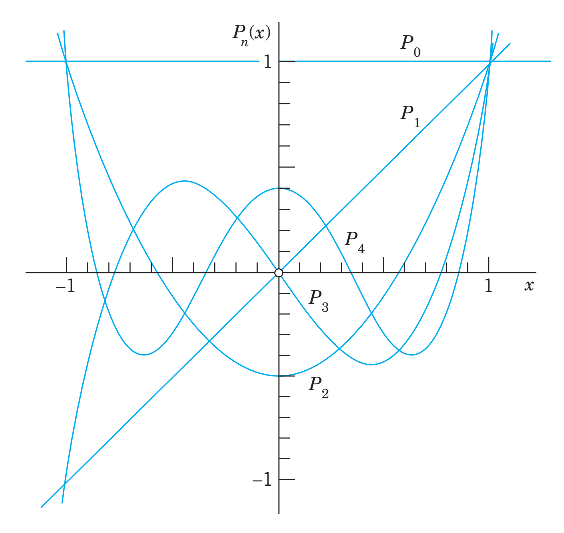


Fig. 107. Legendre polynomials

The Legendre polynomials $P_n(x)$ are **orthogonal** on the interval $-1 \leqq x \leqq 1$, a basic property to be defined and used in making up "Fourier–Legendre series" in the chapter on Fourier series (see Secs. 11.5–11.6).

PROBLEM SET 5.2

1–5 LEGENDRE POLYNOMIALS AND FUNCTIONS

1. **Legendre functions for $n = 0$.** Show that (6) with $n = 0$ gives $P_0(x) = 1$ and (7) gives (use $\ln(1 + x) = x - \frac{1}{2}x^2 + \frac{1}{3}x^3 + \cdots$)

$$y_2(x) = x + \frac{1}{3}x^3 + \frac{1}{5}x^5 + \cdots = \frac{1}{2}\ln\frac{1+x}{1-x}.$$

Verify this by solving (1) with $n = 0$, setting $z = y'$ and separating variables.

2. **Legendre functions for $n = 1$.** Show that (7) with $n = 1$ gives $y_2(x) = P_1(x) = x$ and (6) gives

$$y_1 = 1 - x^2 - \frac{1}{3}x^4 - \frac{1}{5}x^6 - \cdots$$
$$= 1 - \frac{1}{2}x\ln\frac{1+x}{1-x}.$$

3. **Special n.** Derive (11′) from (11).

4. **Legendre's ODE.** Verify that the polynomials in (11′) satisfy (1).

5. Obtain P_6 and P_7.

6–9 CAS PROBLEMS

6. Graph $P_2(x), \cdots, P_{10}(x)$ on common axes. For what x (approximately) and $n = 2, \cdots, 10$ is $|P_n(x)| < \frac{1}{2}$?

7. From what n on will your CAS no longer produce faithful graphs of $P_n(x)$? Why?

8. Graph $Q_0(x), Q_1(x)$, and some further Legendre functions.

9. Substitute $a_s x^s + a_{s+1}x^{s+1} + a_{s+2}x^{s+2}$ into Legendre's equation and obtain the coefficient recursion (4).

10. **TEAM PROJECT. Generating Functions.** Generating functions play a significant role in modern applied mathematics (see [GenRef5]). The idea is simple. If we want to study a certain sequence $(f_n(x))$ and can find a function

$$G(u, x) = \sum_{n=0}^{\infty} f_n(x)u^n,$$

we may obtain properties of $(f_n(x))$ from those of G, which "generates" this sequence and is called a **generating function** of the sequence.

(a) Legendre polynomials. Show that

$$G(u, x) = \frac{1}{\sqrt{1 - 2xu + u^2}} = \sum_{n=0}^{\infty} P_n(x)u^n \qquad (12)$$

is a generating function of the Legendre polynomials. *Hint:* Start from the binomial expansion of $1/\sqrt{1 - v}$, then set $v = 2xu - u^2$, multiply the powers of $2xu - u^2$ out, collect all the terms involving u^n, and verify that the sum of these terms is $P_n(x)u^n$.

(b) Potential theory. Let A_1 and A_2 be two points in space (Fig. 108, $r_2 > 0$). Using (12), show that

$$\frac{1}{r} = \frac{1}{\sqrt{r_1^2 + r_2^2 - 2r_1r_2\cos\theta}}$$
$$= \frac{1}{r_2}\sum_{m=0}^{\infty} P_m(\cos\theta)\left(\frac{r_1}{r_2}\right)^m.$$

This formula has applications in potential theory. (Q/r is the electrostatic potential at A_2 due to a charge Q located at A_1. And the series expresses $1/r$ in terms of the distances of A_1 and A_2 from any origin O and the angle θ between the segments OA_1 and OA_2.)

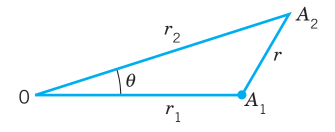


Fig. 108. Team Project 10

(c) Further applications of (12). Show that $P_n(1) = 1$, $P_n(-1) = (-1)^n$, $P_{2n+1}(0) = 0$, and $P_{2n}(0) = (-1)^n \cdot 1 \cdot 3 \cdots (2n - 1)/[2 \cdot 4 \cdots (2n)]$.

11–15 FURTHER FORMULAS

11. **ODE.** Find a solution of $(a^2 - x^2)y'' - 2xy' + n(n + 1)y = 0$, $a \neq 0$, by reduction to the Legendre equation.

12. **Rodrigues's formula (13)**[2] Applying the binomial theorem to $(x^2 - 1)^n$, differentiating it n times term by term, and comparing the result with (11), show that

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}[(x^2 - 1)^n]. \qquad (13)$$

[2]OLINDE RODRIGUES (1794–1851), French mathematician and economist.

13. Rodrigues's formula. Obtain (11′) from (13).

14. Bonnet's recursion.[3] Differentiating (13) with respect to u, using (13) in the resulting formula, and comparing coefficients of u^n, obtain the *Bonnet recursion.*

(14) $(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - np_{n-1}(x),$

where $n = 1, 2, \cdots$. This formula is useful for computations, the loss of significant digits being small (except near zeros). Try (14) out for a few computations of your own choice.

15. Associated Legendre functions $P_n^k(x)$ are needed, e.g., in quantum physics. They are defined by

$$P_n^k(x) = (1-x^2)^{k/2} \frac{d^k p_n(x)}{dx^k} \qquad (15)$$

and are solutions of the ODE

$$(1-x^2)y'' - 2xy' + q(x)y = 0 \qquad (16)$$

where $q(x) = n(n+1) - k^2/(1-x^2)$. Find $P_1^1(x)$, $P_2^1(x)$, $P_2^2(x)$, and $P_4^2(x)$ and verify that they satisfy (16).

5.3 Extended Power Series Method: Frobenius Method

Several second-order ODEs of considerable practical importance—the famous Bessel equation among them—have coefficients that are not analytic (definition in Sec. 5.1), but are "not too bad," so that these ODEs can still be solved by series (power series times a logarithm or times a fractional power of x, etc.). Indeed, the following theorem permits an extension of the power series method. The new method is called the **Frobenius method**.[4] Both methods, that is, the power series method and the Frobenius method, have gained in significance due to the use of software in actual calculations.

THEOREM 1

Frobenius Method

Let $b(x)$ and $c(x)$ be any functions that are analytic at $x = 0$. Then the ODE

$$y'' + \frac{b(x)}{x} y' + \frac{c(x)}{x^2} y = 0 \qquad (1)$$

has at least one solution that can be represented in the form

$$y(x) = x^r \sum_{m=0}^{\infty} a_m x^m = x^r(a_0 + a_1 x + a_2 x^2 + \cdots) \qquad (a_0 \neq 0) \qquad (2)$$

where the exponent r may be any (real or complex) number (and r is chosen so that $a_0 \neq 0$).

The ODE (1) also has a second solution (such that these two solutions are linearly independent) that may be similar to (2) (with a different r and different coefficients) or may contain a logarithmic term. (Details in Theorem 2 below.)

---

[3] OSSIAN BONNET (1819–1892), French mathematician, whose main work was in differential geometry.

[4] GEORG FROBENIUS (1849–1917), German mathematician, professor at ETH Zurich and University of Berlin, student of Karl Weierstrass (see footnote, Sect. 15.5). He is also known for his work on matrices and in group theory.

In this theorem we may replace x by $x - x_0$ with any number x_0. The condition $a_0 \neq 0$ is no restriction; it simply means that we factor out the highest possible power of x.

The singular point of (1) at $x = 0$ is often called a **regular singular point**, a term confusing to the student, which we shall not use.

For example, Bessel's equation (to be discussed in the next section)

$$y'' + \frac{1}{x}y' + \left(\frac{x^2 - \nu^2}{x^2}\right)y = 0 \qquad (\nu \text{ a parameter})$$

is of the form (1) with $b(x) = 1$ and $c(x) = x^2 - \nu^2$ analytic at $x = 0$, so that the theorem applies. This ODE could not be handled in full generality by the power series method.

Similarly, the so-called hypergeometric differential equation (see Problem Set 5.3) also requires the Frobenius method.

The point is that in (2) we have a power series times a single power of x whose exponent r is not restricted to be a nonnegative integer. (The latter restriction would make the whole expression a power series, by definition; see Sec. 5.1.)

The proof of the theorem requires advanced methods of complex analysis and can be found in Ref. [A11] listed in App. 1.

Regular and Singular Points. The following terms are practical and commonly used. A **regular point** of the ODE

$$y'' + p(x)y' + q(x)y = 0$$

is a point x_0 at which the coefficients p and q are analytic. Similarly, a **regular point** of the ODE

$$\widetilde{h}(x)y'' + \widetilde{p}(x)y'(x) + \widetilde{q}(x)y = 0$$

is an x_0 at which $\widetilde{h}, \widetilde{p}, \widetilde{q}$ are analytic and $\widetilde{h}(x_0) \neq 0$ (so what we can divide by $\widetilde{h}$ and get the previous standard form). Then the power series method can be applied. If x_0 is not a regular point, it is called a **singular point**.

Indicial Equation, Indicating the Form of Solutions

We shall now explain the Frobenius method for solving (1). Multiplication of (1) by x^2 gives the more convenient form

$$x^2y'' + xb(x)y' + c(x)y = 0. \qquad \textbf{(1')}$$

We first expand $b(x)$ and $c(x)$ in power series,

$$b(x) = b_0 + b_1x + b_2x^2 + \cdots, \qquad c(x) = c_0 + c_1x + c_2x^2 + \cdots$$

or we do nothing if $b(x)$ and $c(x)$ are polynomials. Then we differentiate (2) term by term, finding

$$\begin{aligned} y'(x) &= \sum_{m=0}^{\infty} (m + r)a_m x^{m+r-1} = x^{r-1}[ra_0 + (r + 1)a_1x + \cdots] \\ y''(x) &= \sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r-2} \\ &= x^{r-2}[r(r - 1)a_0 + (r + 1)ra_1x + \cdots]. \end{aligned} \qquad (2^*)$$

By inserting all these series into (1′) we obtain

$$x^r[r(r-1)a_0 + \cdots] + (b_0 + b_1x + \cdots)x^r(ra_0 + \cdots) \tag{3}$$
$$+ (c_0 + c_1x + \cdots)x^r(a_0 + a_1x + \cdots) = 0.$$

We now equate the sum of the coefficients of each power $x^r, x^{r+1}, x^{r+2}, \cdots$ to zero. This yields a system of equations involving the unknown coefficients a_m. The smallest power is x^r and the corresponding equation is

$$[r(r-1) + b_0r + c_0]a_0 = 0.$$

Since by assumption $a_0 \neq 0$, the expression in the brackets $[\cdots]$ must be zero. This gives

$$r(r-1) + b_0r + c_0 = 0. \tag{4}$$

This important quadratic equation is called the **indicial equation** of the ODE (1). Its role is as follows.

The Frobenius method yields a basis of solutions. One of the two solutions will always be of the form (2), where r is a root of (4). The other solution will be of a form indicated by the indicial equation. There are three cases:

Case 1. Distinct roots not differing by an integer $1, 2, 3, \cdots$.

Case 2. A double root.

Case 3. Roots differing by an integer $1, 2, 3, \cdots$.

Cases 1 and 2 are not unexpected because of the Euler–Cauchy equation (Sec. 2.5), the simplest ODE of the form (1). Case 1 includes complex conjugate roots r_1 and $r_2 = \bar{r}_1$ because $r_1 - r_2 = r_1 - \bar{r}_1 = 2i \operatorname{Im} r_1$ is imaginary, so it cannot be a *real* integer. The form of a basis will be given in Theorem 2 (which is proved in App. 4), without a general theory of convergence, but convergence of the occurring series can be tested in each individual case as usual. Note that in Case 2 *we **must*** have a logarithm, whereas in Case 3 we *may* or *may not*.

THEOREM 2

Frobenius Method. Basis of Solutions. Three Cases

Suppose that the ODE (1) *satisfies the assumptions in Theorem* 1. *Let* r_1 *and* r_2 *be the roots of the indicial equation* (4). *Then we have the following three cases.*

Case* 1. *Distinct Roots Not Differing by an Integer. *A basis is*

$$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdots) \tag{5}$$

and

$$y_2(x) = x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots) \tag{6}$$

with coefficients obtained successively from (3) *with* $r = r_1$ *and* $r = r_2$, *respectively.*

Case 2. Double Root $r_1 = r_2 = r$. *A basis is*

(7) $$y_1(x) = x^r(a_0 + a_1x + a_2x^2 + \cdots) \qquad [r = \tfrac{1}{2}(1 - b_0)]$$

(*of the same general form as before*) *and*

(8) $$y_2(x) = y_1(x) \ln x + x^r(A_1x + A_2x^2 + \cdots) \qquad (x > 0).$$

Case 3. Roots Differing by an Integer. *A basis is*

(9) $$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdots)$$

(*of the same general form as before*) *and*

(10) $$y_2(x) = ky_1(x) \ln x + x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots),$$

where the roots are so denoted that $r_1 - r_2 > 0$ and k may turn out to be zero.

Typical Applications

Technically, the Frobenius method is similar to the power series method, once the roots of the indicial equation have been determined. However, (5)–(10) merely indicate the general form of a basis, and a second solution can often be obtained more rapidly by reduction of order (Sec. 2.1).

EXAMPLE 1 Euler–Cauchy Equation, Illustrating Cases 1 and 2 and Case 3 without a Logarithm

For the Euler–Cauchy equation (Sec. 2.5)

$$x^2y'' + b_0xy' + c_0y = 0 \qquad (b_0, c_0 \text{ constant})$$

substitution of $y = x^r$ gives the auxiliary equation

$$r(r - 1) + b_0r + c_0 = 0,$$

which is the indicial equation [and $y = x^r$ is a very special form of (2)!]. For different roots r_1, r_2 we get a basis $y_1 = x^{r_1}$, $y_2 = x^{r_2}$, and for a double root r we get a basis x^r, $x^r \ln x$. Accordingly, for this simple ODE, Case 3 plays no extra role. ■

EXAMPLE 2 Illustration of Case 2 (Double Root)

Solve the ODE

(11) $$x(x - 1)y'' + (3x - 1)y' + y = 0.$$

(This is a special hypergeometric equation, as we shall see in the problem set.)

Solution. Writing (11) in the standard form (1), we see that it satisfies the assumptions in Theorem 1. [What are $b(x)$ and $c(x)$ in (11)?] By inserting (2) and its derivatives (2*) into (11) we obtain

(12) $$\sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r} - \sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r-1}$$
$$+ 3\sum_{m=0}^{\infty} (m + r)a_m x^{m+r} - \sum_{m=0}^{\infty} (m + r)a_m x^{m+r-1} + \sum_{m=0}^{\infty} a_m x^{m+r} = 0.$$

The smallest power is x^{r-1}, occurring in the second and the fourth series; by equating the sum of its coefficients to zero we have

$$[-r(r-1)-r]a_0 = 0, \qquad \text{thus} \qquad r^2 = 0.$$

Hence this indicial equation has the double root $r = 0$.

First Solution. We insert this value $r = 0$ into (12) and equate the sum of the coefficients of the power x^s to zero, obtaining

$$s(s-1)a_s - (s+1)sa_{s+1} + 3sa_s - (s+1)a_{s+1} + a_s = 0$$

thus $a_{s+1} = a_s$. Hence $a_0 = a_1 = a_2 = \cdots$, and by choosing $a_0 = 1$ we obtain the solution

$$y_1(x) = \sum_{m=0}^{\infty} x^m = \frac{1}{1-x} \qquad (|x| < 1).$$

Second Solution. We get a second independent solution y_2 by the method of reduction of order (Sec. 2.1), substituting $y_2 = uy_1$ and its derivatives into the equation. This leads to (9), Sec. 2.1, which we shall use in this example, instead of starting reduction of order from scratch (as we shall do in the next example). In (9) of Sec. 2.1 we have $p = (3x-1)/(x^2-x)$, the coefficient of y' in (11) ***in standard form***. By partial fractions,

$$-\int p\,dx = -\int \frac{3x-1}{x(x-1)}\,dx = -\int\left(\frac{2}{x-1} + \frac{1}{x}\right)dx = -2\ln(x-1) - \ln x.$$

Hence (9), Sec. 2.1, becomes

$$u' = U = y_1^{-2}e^{-\int p\,dx} = \frac{(x-1)^2}{(x-1)^2 x} = \frac{1}{x}, \qquad u = \ln x, \qquad y_2 = uy_1 = \frac{\ln x}{1-x}.$$

y_1 and y_2 are shown in Fig. 109. These functions are linearly independent and thus form a basis on the interval $0 < x < 1$ (as well as on $1 < x < \infty$). ■

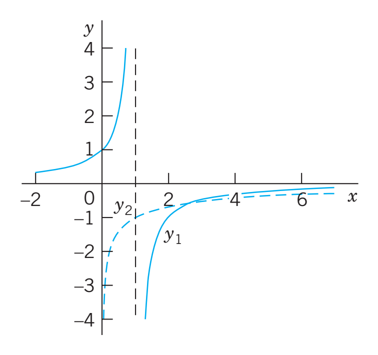


Fig. 109. Solutions in Example 2

EXAMPLE 3 Case 3, Second Solution with Logarithmic Term

Solve the ODE

$$(x^2 - x)y'' - xy' + y = 0. \tag{13}$$

Solution. Substituting (2) and (2*) into (13), we have

$$(x^2 - x)\sum_{m=0}^{\infty}(m+r)(m+r-1)a_m x^{m+r-2} - x\sum_{m=0}^{\infty}(m+r)a_m x^{m+r-1} + \sum_{m=0}^{\infty} a_m x^{m+r} = 0.$$

We now take x^2, x, and x inside the summations and collect all terms with power x^{m+r} and simplify algebraically,

$$\sum_{m=0}^{\infty} (m+r-1)^2 a_m x^{m+r} - \sum_{m=0}^{\infty} (m+r)(m+r-1) a_m x^{m+r-1} = 0.$$

In the first series we set $m = s$ and in the second $m = s + 1$, thus $s = m - 1$. Then

$$\sum_{s=0}^{\infty} (s+r-1)^2 a_s x^{s+r} - \sum_{s=-1}^{\infty} (s+r+1)(s+r) a_{s+1} x^{s+r} = 0. \tag{14}$$

The lowest power is x^{r-1} (take $s = -1$ in the second series) and gives the indicial equation

$$r(r-1) = 0.$$

The roots are $r_1 = 1$ and $r_2 = 0$. They differ by an integer. This is Case 3.

First Solution. From (14) with $r = r_1 = 1$ we have

$$\sum_{s=0}^{\infty} [s^2 a_s - (s+2)(s+1) a_{s+1}] x^{s+1} = 0.$$

This gives the recurrence relation

$$a_{s+1} = \frac{s^2}{(s+2)(s+1)} a_s \qquad (s = 0, 1, \cdots).$$

Hence $a_1 = 0, a_2 = 0, \cdots$ successively. Taking $a_0 = 1$, we get as a first solution $y_1 = x^{r_1} a_0 = x$.

Second Solution. Applying reduction of order (Sec. 2.1), we substitute $y_2 = y_1 u = xu$, $y_2' = xu' + u$ and $y_2'' = xu'' + 2u'$ into the ODE, obtaining

$$(x^2 - x)(xu'' + 2u') - x(xu' + u) + xu = 0.$$

xu drops out. Division by x and simplification give

$$(x^2 - x)u'' + (x - 2)u' = 0.$$

From this, using partial fractions and integrating (taking the integration constant zero), we get

$$\frac{u''}{u'} = -\frac{x-2}{x^2-x} = -\frac{2}{x} + \frac{1}{1-x}, \qquad \ln u' = \ln\left|\frac{x-1}{x^2}\right|.$$

Taking exponents and integrating (again taking the integration constant zero), we obtain

$$u' = \frac{x-1}{x^2} = \frac{1}{x} - \frac{1}{x^2}, \qquad u = \ln x + \frac{1}{x}, \qquad y_2 = xu = x \ln x + 1.$$

y_1 and y_2 are linearly independent, and y_2 has a logarithmic term. Hence y_1 and y_2 constitute a basis of solutions for positive x. ■

The Frobenius method solves the **hypergeometric equation**, whose solutions include many known functions as special cases (see the problem set). In the next section we use the method for solving Bessel's equation.

PROBLEM SET 5.3

1. WRITING PROJECT. Power Series Method and Frobenius Method. Write a report of 2–3 pages explaining the difference between the two methods. No proofs. Give simple examples of your own.

2–13 FROBENIUS METHOD

Find a basis of solutions by the Frobenius method. Try to identify the series as expansions of known functions. Show the details of your work.

2. $(x + 2)^2y'' + (x + 2)y' - y = 0$

3. $xy'' + 2y' + xy = 0$

4. $xy'' + y = 0$

5. $xy'' + (2x + 1)y' + (x + 1)y = 0$

6. $xy'' + 2x^3y' + (x^2 - 2)y = 0$

7. $y'' + (x - 1)y = 0$

8. $xy'' + y' - xy = 0$

9. $2x(x - 1)y'' - (x + 1)y' + y = 0$

10. $xy'' + 2y' + 4xy = 0$

11. $xy'' + (2 - 2x)y' + (x - 2)y = 0$

12. $x^2y'' + 6xy' + (4x^2 + 6)y = 0$

13. $xy'' + (1 - 2x)y' + (x - 1)y = 0$

14. TEAM PROJECT. Hypergeometric Equation, Series, and Function. Gauss's hypergeometric ODE[5] is

$$x(1 - x)y'' + [c - (a + b + 1)x]y' - aby = 0. \tag{15}$$

Here, a, b, c are constants. This ODE is of the form $p_2y'' + p_1y' + p_0y = 0$, where p_2, p_1, p_0 are polynomials of degree 2, 1, 0, respectively. These polynomials are written so that the series solution takes a most practical form, namely,

$$y_1(x) = 1 + \frac{ab}{1!\,c}x + \frac{a(a + 1)b(b + 1)}{2!\,c(c + 1)}x^2 + \frac{a(a + 1)(a + 2)b(b + 1)(b + 2)}{3!\,c(c + 1)(c + 2)}x^3 + \cdots. \tag{16}$$

This series is called the **hypergeometric series**. Its sum $y_1(x)$ is called the **hypergeometric function** and is denoted by $F(a, b, c; x)$. Here, $c \neq 0, -1, -2, \cdots$. By choosing specific values of a, b, c we can obtain an incredibly large number of special functions as solutions of (15) [see the small sample of elementary functions in part (c)]. This accounts for the importance of (15).

(a) Hypergeometric series and function. Show that the indicial equation of (15) has the roots $r_1 = 0$ and $r_2 = 1 - c$. Show that for $r_1 = 0$ the Frobenius method gives (16). Motivate the name for (16) by showing that

$$F(1, 1, 1; x) = F(1, b, b; x) = F(a, 1, a; x) = \frac{1}{1 - x}.$$

(b) Convergence. For what a or b will (16) reduce to a polynomial? Show that for any other a, b, c $(c \neq 0, -1, -2, \cdots)$ the series (16) converges when $|x| < 1$.

(c) Special cases. Show that

$$(1 + x)^n = F(-n, b, b; -x),$$
$$(1 - x)^n = 1 - nxF(1 - n, 1, 2; x),$$
$$\arctan x = xF(\tfrac{1}{2}, 1, \tfrac{3}{2}; -x^2)$$
$$\arcsin x = xF(\tfrac{1}{2}, \tfrac{1}{2}, \tfrac{3}{2}; x^2),$$
$$\ln(1 + x) = xF(1, 1, 2; -x),$$
$$\ln\frac{1 + x}{1 - x} = 2xF(\tfrac{1}{2}, 1, \tfrac{3}{2}; x^2).$$

Find more such relations from the literature on special functions, for instance, from [GenRef1] in App. 1.

(d) Second solution. Show that for $r_2 = 1 - c$ the Frobenius method yields the following solution (where $c \neq 2, 3, 4, \cdots$):

$$y_2(x) = x^{1-c}\left(1 + \frac{(a - c + 1)(b - c + 1)}{1!\,(-c + 2)}x + \frac{(a - c + 1)(a - c + 2)(b - c + 1)(b - c + 2)}{2!\,(-c + 2)(-c + 3)}x^2 + \cdots\right). \tag{17}$$

Show that

$$y_2(x) = x^{1-c}F(a - c + 1, b - c + 1, 2 - c; x).$$

(e) On the generality of the hypergeometric equation. Show that

$$(t^2 + At + B)\ddot{y} + (Ct + D)\dot{y} + Ky = 0 \tag{18}$$

[5]CARL FRIEDRICH GAUSS (1777–1855), great German mathematician. He already made the first of his great discoveries as a student at Helmstedt and Göttingen. In 1807 he became a professor and director of the Observatory at Göttingen. His work was of basic importance in algebra, number theory, differential equations, differential geometry, non-Euclidean geometry, complex analysis, numeric analysis, astronomy, geodesy, electromagnetism, and theoretical mechanics. He also paved the way for a general and systematic use of complex numbers.

with $\dot{y} = dy/dt$, etc., constant A, B, C, D, K, and $t^2 + At + B = (t - t_1)(t - t_2)$, $t_1 \neq t_2$, can be reduced to the hypergeometric equation with independent variable

$$x = \frac{t - t_1}{t_2 - t_1}$$

and parameters related by $Ct_1 + D = -c(t_2 - t_1)$, $C = a + b + 1$, $K = ab$. From this you see that (15) is a "normalized form" of the more general (18) and that various cases of (18) can thus be solved in terms of hypergeometric functions.

15–20 HYPERGEOMETRIC ODE

Find a general solution in terms of hypergeometric functions.

15. $2x(1 - x)y'' - (1 + 6x)y' - 2y = 0$

16. $x(1 - x)y'' + (\frac{1}{2} + 2x)y' - 2y = 0$

17. $4x(1 - x)y'' + y' + 8y = 0$

18. $4(t^2 - 3t + 2)\ddot{y} - 2\dot{y} + y = 0$

19. $2(t^2 - 5t + 6)\ddot{y} + (2t - 3)\dot{y} - 8y = 0$

20. $3t(1 + t)\ddot{y} + t\dot{y} - y = 0$

5.4 Bessel's Equation. Bessel Functions $J_\nu(x)$

One of the most important ODEs in applied mathematics in **Bessel's equation**,[6]

$$x^2y'' + xy' + (x^2 - \nu^2)y = 0 \tag{1}$$

where the parameter ν (nu) is a given real number which is positive or zero. Bessel's equation often appears if a problem shows cylindrical symmetry, for example, as the membranes in Sec.12.9. The equation satisfies the assumptions of Theorem 1. To see this, divide (1) by x^2 to get the standard form $y'' + y'/x + (1 - \nu^2/x^2)y = 0$. Hence, according to the Frobenius theory, it has a solution of the form

$$y(x) = \sum_{m=0}^{\infty} a_m x^{m+r} \qquad (a_0 \neq 0). \tag{2}$$

Substituting (2) and its first and second derivatives into Bessel's equation, we obtain

$$\sum_{m=0}^{\infty} (m + r)(m + r - 1)a_m x^{m+r} + \sum_{m=0}^{\infty} (m + r)a_m x^{m+r}$$
$$+ \sum_{m=0}^{\infty} a_m x^{m+r+2} - \nu^2 \sum_{m=0}^{\infty} a_m x^{m+r} = 0.$$

We equate the sum of the coefficients of x^{s+r} to zero. Note that this power x^{s+r} corresponds to $m = s$ in the first, second, and fourth series, and to $m = s - 2$ in the third series. Hence for $s = 0$ and $s = 1$, the third series does not contribute since $m \geqq 0$.

[6]FRIEDRICH WILHELM BESSEL (1784–1846), German astronomer and mathematician, studied astronomy on his own in his spare time as an apprentice of a trade company and finally became director of the new Königsberg Observatory.

Formulas on Bessel functions are contained in Ref. [GenRef10] and the standard treatise [A13].

For $s = 2, 3, \cdots$ all four series contribute, so that we get a general formula for all these s. We find

$$\text{(3)}\quad \begin{aligned} &\text{(a)} && r(r-1)a_0 + ra_0 - \nu^2 a_0 = 0 && (s = 0) \\ &\text{(b)} && (r+1)ra_1 + (r+1)a_1 - \nu^2 a_1 = 0 && (s = 1) \\ &\text{(c)} && (s+r)(s+r-1)a_s + (s+r)a_s + a_{s-2} - \nu^2 a_s = 0 && (s = 2, 3, \cdots). \end{aligned}$$

From (3a) we obtain the **indicial equation** by dropping a_0,

$$\textbf{(4)}\qquad (r+\nu)(r-\nu) = 0.$$

The roots are $r_1 = \nu\ (\geqq 0)$ and $r_2 = -\nu$.

Coefficient Recursion for $r = r_1 = \nu$. For $r = \nu$, Eq. (3b) reduces to $(2\nu + 1)a_1 = 0$. Hence $a_1 = 0$ since $\nu \geqq 0$. Substituting $r = \nu$ in (3c) and combining the three terms containing a_s gives simply

$$\text{(5)}\qquad (s + 2\nu)sa_s + a_{s-2} = 0.$$

Since $a_1 = 0$ and $\nu \geqq 0$, it follows from (5) that $a_3 = 0, a_5 = 0, \cdots$. Hence we have to deal only with *even-numbered* coefficients a_s with $s = 2m$. For $s = 2m$, Eq. (5) becomes

$$(2m + 2\nu)2ma_{2m} + a_{2m-2} = 0.$$

Solving for a_{2m} gives the recursion formula

$$\textbf{(6)}\qquad a_{2m} = -\frac{1}{2^2 m(\nu + m)}\, a_{2m-2}, \qquad m = 1, 2, \cdots.$$

From (6) we can now determine $a_2, a_4, \cdots$ successively. This gives

$$a_2 = -\frac{a_0}{2^2(\nu+1)}$$

$$a_4 = -\frac{a_2}{2^2 2(\nu+2)} = \frac{a_0}{2^4 2!\,(\nu+1)(\nu+2)}$$

and so on, and in general

$$\textbf{(7)}\qquad a_{2m} = \frac{(-1)^m a_0}{2^{2m} m!\,(\nu+1)(\nu+2)\cdots(\nu+m)}, \qquad m = 1, 2, \cdots.$$

Bessel Functions $J_n(x)$ for Integer $\nu = n$

Integer values of ν are denoted by n. This is standard. For $\nu = n$ the relation (7) becomes

$$\text{(8)}\qquad a_{2m} = \frac{(-1)^m a_0}{2^{2m} m!\,(n+1)(n+2)\cdots(n+m)}, \qquad m = 1, 2, \cdots.$$

a_0 is still arbitrary, so that the series (2) with these coefficients would contain this arbitrary factor a_0. This would be a highly impractical situation for developing formulas or computing values of this new function. Accordingly, we have to make a choice. The choice $a_0 = 1$ would be possible. A simpler series (2) could be obtained if we could absorb the growing product $(n+1)(n+2)\cdots(n+m)$ into a factorial function $(n+m)!$ What should be our choice? Our choice should be

$$a_0 = \frac{1}{2^n n!} \tag{9}$$

because then $n!\,(n+1)\cdots(n+m) = (n+m)!$ in (8), so that (8) simply becomes

$$a_{2m} = \frac{(-1)^m}{2^{2m+n} m!\,(n+m)!}, \qquad m = 1, 2, \cdots. \tag{10}$$

By inserting these coefficients into (2) and remembering that $c_1 = 0, c_3 = 0, \cdots$ we obtain a particular solution of Bessel's equation that is denoted by $J_n(x)$:

$$J_n(x) = x^n \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+n} m!\,(n+m)!} \qquad (n \geqq 0). \tag{11}$$

$J_n(x)$ is called the **Bessel function of the first kind** *of order n*. The series (11) converges for all x, as the ratio test shows. Hence $J_n(x)$ is defined for all x. The series converges very rapidly because of the factorials in the denominator.

EXAMPLE 1 Bessel Functions $J_0(x)$ and $J_1(x)$

For $n = 0$ we obtain from (11) the **Bessel function of order 0**

$$J_0(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m}(m!)^2} = 1 - \frac{x^2}{2^2(1!)^2} + \frac{x^4}{2^4(2!)^2} - \frac{x^6}{2^6(3!)^2} + - \cdots \tag{12}$$

which looks similar to a cosine (Fig. 110). For $n = 1$ we obtain the **Bessel function of order 1**

$$J_1(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{2^{2m+1} m!\,(m+1)!} = \frac{x}{2} - \frac{x^3}{2^3 1!\,2!} + \frac{x^5}{2^5 2!\,3!} - \frac{x^7}{2^7 3!\,4!} + - \cdots, \tag{13}$$

which looks similar to a sine (Fig. 110). But the zeros of these functions are not completely regularly spaced (see also Table A1 in App. 5) and the height of the "waves" decreases with increasing x. Heuristically, n^2/x^2 in (1) in standard form [(1) divided by x^2] is zero (if $n = 0$) or small in absolute value for large x, and so is y'/x, so that then Bessel's equation comes close to $y'' + y = 0$, the equation of $\cos x$ and $\sin x$; also y'/x acts as a "damping term," in part responsible for the decrease in height. One can show that for large x,

$$J_n(x) \sim \sqrt{\frac{2}{\pi x}} \cos\left(x - \frac{n\pi}{2} - \frac{\pi}{4}\right) \tag{14}$$

where $\sim$ is read "**asymptotically equal**" and means that *for fixed n* the quotient of the two sides approaches 1 as $x \to \infty$.

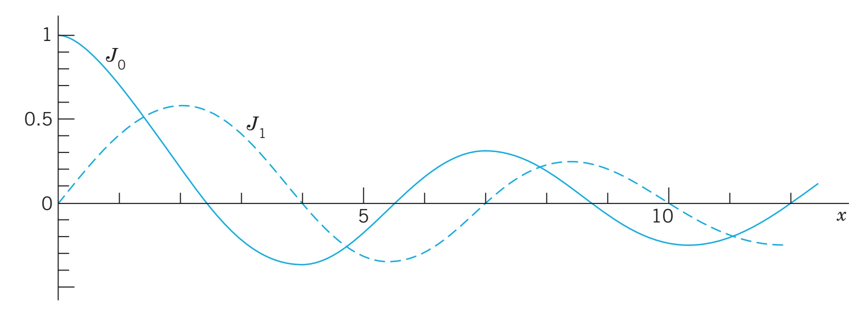


Fig. 110. Bessel functions of the first kind J_0 and J_1

Formula (14) is surprisingly accurate even for smaller $x\,(>0)$. For instance, it will give you good starting values in a computer program for the basic task of computing zeros. For example, for the first three zeros of J_0 you obtain the values 2.356 (2.405 exact to 3 decimals, error 0.049), 5.498 (5.520, error 0.022), 8.639 (8.654, error 0.015), etc. ■

Bessel Functions $J_\nu(x)$ for any $\nu \geqq 0$. Gamma Function

We now proceed from integer $\nu = n$ to any $\nu \geqq 0$. We had $a_0 = 1/(2^n n!)$ in (9). So we have to extend the factorial function $n!$ to any $\nu \geqq 0$. For this we choose

$$a_0 = \frac{1}{2^\nu \Gamma(\nu + 1)} \tag{15}$$

with the **gamma function** $\Gamma(\nu + 1)$ defined by

$$\Gamma(\nu + 1) = \int_0^\infty e^{-t} t^\nu \, dt \qquad (\nu > -1). \tag{16}$$

(**CAUTION!** Note the convention $\nu + 1$ on the left but ν in the integral.) Integration by parts gives

$$\Gamma(\nu + 1) = -e^{-t} t^\nu \Big|_0^\infty + \nu \int_0^\infty e^{-t} t^{\nu-1} \, dt = 0 + \nu\Gamma(\nu).$$

This is the basic functional relation of the gamma function

$$\Gamma(\nu + 1) = \nu\Gamma(\nu). \tag{17}$$

Now from (16) with $\nu = 0$ and then by (17) we obtain

$$\Gamma(1) = \int_0^\infty e^{-t} \, dt = -e^{-t} \Big|_0^\infty = 0 - (-1) = 1$$

and then $\Gamma(2) = 1 \cdot \Gamma(1) = 1!$, $\Gamma(3) = 2\Gamma(1) = 2!$ and in general

$$\Gamma(n + 1) = n! \qquad (n = 0, 1, \cdots). \tag{18}$$

Hence ***the gamma function generalizes the factorial function* to arbitrary positive** ν. Thus (15) with $\nu = n$ agrees with (9).

Furthermore, from (7) with a_0 given by (15) we first have

$$a_{2m} = \frac{(-1)^m}{2^{2m} m!\,(\nu+1)(\nu+2)\cdots(\nu+m)2^\nu\Gamma(\nu+1)}.$$

Now (17) gives $(\nu+1)\Gamma(\nu+1) = \Gamma(\nu+2)$, $\quad(\nu+2)\Gamma(\nu+2) = \Gamma(\nu+3)$ and so on, so that

$$(\nu+1)(\nu+2)\cdots(\nu+m)\Gamma(\nu+1) = \Gamma(\nu+m+1).$$

Hence because of our (standard!) choice (15) of a_0 the coefficients (7) are simply

$$a_{2m} = \frac{(-1)^m}{2^{2m+\nu} m!\,\Gamma(\nu+m+1)}. \tag{19}$$

With these coefficients and $r = r_1 = \nu$ we get from (2) a particular solution of (1), denoted by $J_\nu(x)$ and given by

(20)
$$J_\nu(x) = x^\nu \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+\nu} m!\,\Gamma(\nu+m+1)}.$$

$J_\nu(x)$ is called the **Bessel function of the first kind of order** $\boldsymbol{\nu}$. The series (20) converges for all x, as one can verify by the ratio test.

Discovery of Properties from Series

Bessel functions are a model case for showing how to discover properties and relations of functions from series by which they are *defined.* Bessel functions satisfy an incredibly large number of relationships—look at Ref. [A13] in App. 1; also, find out what your CAS knows. In Theorem 3 we shall discuss four formulas that are backbones in applications and theory.

THEOREM 1

Derivatives, Recursions

The derivative of $J_\nu(x)$ with respect to x can be expressed by $J_{\nu-1}(x)$ or $J_{\nu+1}(x)$ by the formulas

(21)
(a) $$[x^\nu J_\nu(x)]' = x^\nu J_{\nu-1}(x)$$
(b) $$[x^{-\nu} J_\nu(x)]' = -x^{-\nu} J_{\nu+1}(x).$$

Furthermore, $J_\nu(x)$ and its derivative satisfy the recurrence relations

(21)
(c) $$J_{\nu-1}(x) + J_{\nu+1}(x) = \frac{2\nu}{x} J_\nu(x)$$
(d) $$J_{\nu-1}(x) - J_{\nu+1}(x) = 2J'_\nu(x).$$

PROOF **(a)** We multiply (20) by x^ν and take $x^{2\nu}$ under the summation sign. Then we have

$$x^\nu J_\nu(x) = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+2\nu}}{2^{2m+\nu} m!\, \Gamma(\nu + m + 1)}.$$

We now differentiate this, cancel a factor 2, pull $x^{2\nu-1}$ out, and use the functional relationship $\Gamma(\nu + m + 1) = (\nu + m)\Gamma(\nu + m)$ [see (17)]. Then (20) with $\nu - 1$ instead of ν shows that we obtain the right side of (21a). Indeed,

$$(x^\nu J_\nu)' = \sum_{m=0}^{\infty} \frac{(-1)^m 2(m + \nu)x^{2m+2\nu-1}}{2^{2m+\nu} m!\, \Gamma(\nu + m + 1)} = x^\nu x^{\nu-1} \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{2^{2m+\nu-1} m!\, \Gamma(\nu + m)}.$$

(b) Similarly, we multiply (20) by $x^{-\nu}$, so that x^ν in (20) cancels. Then we differentiate, cancel $2m$, and use $m! = m(m - 1)!$. This gives, with $m = s + 1$,

$$(x^{-\nu} J_\nu)' = \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m+\nu-1}(m - 1)!\, \Gamma(\nu + m + 1)} = \sum_{s=0}^{\infty} \frac{(-1)^{s+1} x^{2s+1}}{2^{2s+\nu+1} s!\, \Gamma(\nu + s + 2)}.$$

Equation (20) with $\nu + 1$ instead of ν and s instead of m shows that the expression on the right is $-x^{-\nu} J_{\nu+1}(x)$. This proves (21b).

(c), **(d)** We perform the differentiation in (21a). Then we do the same in (21b) and multiply the result on both sides by $x^{2\nu}$. This gives

$$\text{(a*)} \qquad \nu x^{\nu-1} J_\nu + x^\nu J_\nu' = x^\nu J_{\nu-1}$$

$$\text{(b*)} \qquad -\nu x^{\nu-1} J_\nu + x^\nu J_\nu' = -x^\nu J_{\nu+1}.$$

Substracting (b*) from (a*) and dividing the result by x^ν gives (21c). Adding (a*) and (b*) and dividing the result by x^ν gives (21d). ■

EXAMPLE 2 Application of Theorem 1 in Evaluation and Integration

Formula (21c) can be used recursively in the form

$$J_{\nu+1}(x) = \frac{2\nu}{x} J_\nu(x) - J_{\nu-1}(x)$$

for calculating Bessel functions of higher order from those of lower order. For instance, $J_2(x) = 2J_1(x)/x - J_0(x)$, so that J_2 can be obtained from tables of J_0 and J_1 (in App. 5 or, more accurately, in Ref. [GenRef1] in App. 1).

To illustrate how Theorem 1 helps in integration, we use (21b) with $\nu = 3$ integrated on both sides. This evaluates, for instance, the integral

$$I = \int_1^2 x^{-3} J_4(x)\, dx = -x^{-3} J_3(x) \Big|_1^2 = -\frac{1}{8} J_3(2) + J_3(1).$$

A table of J_3 (on p. 398 of Ref. [GenRef1]) or your CAS will give you

$$-\tfrac{1}{8} \cdot 0.128943 + 0.019563 = 0.003445.$$

Your CAS (or a human computer in precomputer times) obtains J_3 from (21), first using (21c) with $\nu = 2$, that is, $J_3 = 4x^{-1} J_2 - J_1$, then (21c) with $\nu = 1$, that is, $J_2 = 2x^{-1} J_1 - J_0$. Together,

$$\begin{aligned} I &= x^{-3}(4x^{-1}(2x^{-1}J_1 - J_0) - J_1)\Big|_1^2 \\ &= -\tfrac{1}{8}[2J_1(2) - 2J_0(2) - J_1(2)] + [8J_1(1) - 4J_0(1) - J_1(1)] \\ &= -\tfrac{1}{8}J_1(2) + \tfrac{1}{4}J_0(2) + 7J_1(1) - 4J_0(1). \end{aligned}$$

This is what you get, for instance, with Maple if you type int$(\cdots)$. And if you type evalf(int$(\cdots)$), you obtain 0.003445448, in agreement with the result near the beginning of the example. ■

Bessel Functions J_ν with Half-Integer ν Are Elementary

We discover this remarkable fact as another property obtained from the series (20) and confirm it in the problem set by using Bessel's ODE.

EXAMPLE 3 Elementary Bessel Functions J_ν with $\nu = \pm\frac{1}{2}, \pm\frac{3}{2}, \pm\frac{5}{2}, \cdots$. The Value $\Gamma(\frac{1}{2})$

We first prove (Fig. 111)

(22) $$\text{(a)}\quad J_{1/2}(x) = \sqrt{\frac{2}{\pi x}}\sin x, \qquad \text{(b)}\quad J_{-1/2}(x) = \sqrt{\frac{2}{\pi x}}\cos x.$$

The series (20) with $\nu = \frac{1}{2}$ is

$$J_{1/2}(x) = \sqrt{x}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m}}{2^{2m+1/2}m!\,\Gamma(m+\frac{3}{2})} = \sqrt{\frac{2}{x}}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+1}}{2^{2m+1}m!\,\Gamma(m+\frac{3}{2})}.$$

The denominator can be written as a product AB, where (use (16) in B)

$$\begin{aligned} A &= 2^m m! = 2m(2m-2)(2m-4)\cdots 4\cdot 2, \\ B &= 2^{m+1}\Gamma(m+\tfrac{3}{2}) = 2^{m+1}(m+\tfrac{1}{2})(m-\tfrac{1}{2})\cdots\tfrac{3}{2}\cdot\tfrac{1}{2}\Gamma(\tfrac{1}{2}) \\ &= (2m+1)(2m-1)\cdots 3\cdot 1\cdot\sqrt{\pi}; \end{aligned}$$

here we used (proof below)

(23) $$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}.$$

The product of the right sides of A and B can be written

$$AB = (2m+1)2m(2m-1)\cdots 3\cdot 2\cdot 1\sqrt{\pi} = (2m+1)!\sqrt{\pi}.$$

Hence

$$J_{1/2}(x) - \sqrt{\frac{2}{\pi x}}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m+1}}{(2m+1)!} = \sqrt{\frac{2}{\pi x}}\sin x.$$

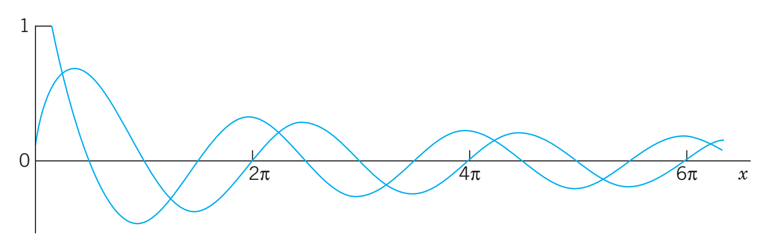


Fig. 111. Bessel functions $J_{1/2}$ and $J_{-1/2}$

This proves (22a). Differentiation and the use of (21a) with $\nu = \frac{1}{2}$ now gives

$$[\sqrt{x}J_{1/2}(x)]' = \sqrt{\frac{2}{\pi}}\cos x = x^{1/2}J_{-1/2}(x).$$

This proves (22b). From (22) follow further formulas successively by (21c), used as in Example 2.

We finally prove $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ by a standard trick worth remembering. In (15) we set $t = u^2$. Then $dt = 2u\,du$ and

$$\Gamma\left(\frac{1}{2}\right) = \int_0^\infty e^{-t}t^{-1/2}\,dt = 2\int_0^\infty e^{-u^2}\,du.$$

We square on both sides, write v instead of u in the second integral, and then write the product of the integrals as a double integral:

$$\Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^\infty e^{-u^2}\,du \int_0^\infty e^{-v^2}\,dv = 4\int_0^\infty\int_0^\infty e^{-(u^2+v^2)}\,du\,dv.$$

We now use polar coordinates r, θ by setting $u = r\cos\theta$, $v = r\sin\theta$. Then the element of area is $du\,dv = r\,dr\,d\theta$ and we have to integrate over r from 0 to ∞ and over θ from 0 to $\pi/2$ (that is, over the first quadrant of the uv-plane):

$$\Gamma\left(\frac{1}{2}\right)^2 = 4\int_0^{\pi/2}\int_0^\infty e^{-r^2}r\,dr\,d\theta = 4\cdot\frac{\pi}{2}\int_0^\infty e^{-r^2}r\,dr = 2\left(-\frac{1}{2}\right)e^{-r^2}\Big|_0^\infty = \pi.$$

By taking the square root on both sides we obtain (23). ■

General Solution. Linear Dependence

For a general solution of Bessel's equation (1) in addition to J_ν we need a second linearly independent solution. For ν not an integer this is easy. Replacing ν by $-\nu$ in (20), we have

$$J_{-\nu}(x) = x^{-\nu}\sum_{m=0}^{\infty}\frac{(-1)^m x^{2m}}{2^{2m-\nu}m!\,\Gamma(m-\nu+1)}. \tag{24}$$

Since Bessel's equation involves ν^2, the functions J_ν and $J_{-\nu}$ are solutions of the equation for the same ν. If ν is not an integer, they are linearly independent, because the first terms in (20) and in (24) are finite nonzero multiples of x^ν and $x^{-\nu}$. Thus, if ν is not an integer, a general solution of Bessel's equation for all $x \neq 0$ is

$$y(x) = c_1 J_\nu(x) + c_2 J_{-\nu}(x)$$

This cannot be the general solution for an integer $\nu = n$ because, in that case, we have linear dependence. It can be seen that the first terms in (20) and (24) are finite nonzero multiples of x^ν and $x^{-\nu}$, respectively. This means that, for any integer $\nu = n$, we have linear dependence because

$$J_{-n}(x) = (-1)^n J_n(x) \qquad (n = 1, 2, \cdots). \tag{25}$$

PROOF To prove (25), we use (24) and let ν approach a positive integer n. Then the gamma function in the coefficients of the first n terms becomes infinite (see Fig. 553 in App. A3.1), the coefficients become zero, and the summation starts with $m = n$. Since in this case $\Gamma(m - n + 1) = (m - n)!$ by (18), we obtain

$$(26) \qquad J_{-n}(x) = \sum_{m=n}^{\infty} \frac{(-1)^m x^{2m-n}}{2^{2m-n} m!\,(m-n)!} = \sum_{s=0}^{\infty} \frac{(-1)^{n+s} x^{2s+n}}{2^{2s+n}(n+s)!\,s!} \qquad (m = n + s).$$

The last series represents $(-1)^n J_n(x)$, as you can see from (11) with m replaced by s. This completes the proof. ■

The difficulty caused by (25) will be overcome in the next section by introducing further Bessel functions, called *of the second kind* and denoted by Y_ν.

PROBLEM SET 5.4

1. Convergence. Show that the series (11) converges for all x. Why is the convergence very rapid?

2–10 ODEs REDUCIBLE TO BESSEL'S ODE

This is just a sample of such ODEs; some more follow in the next problem set. Find a general solution in terms of J_ν and $J_{-\nu}$ or indicate when this is not possible. Use the indicated substitutions. Show the details of your work.

2. $x^2y'' + xy' + (x^2 - \frac{4}{49})y = 0$

3. $xy'' + y' + \frac{1}{4}y = 0 \quad (\sqrt{x} = z)$

4. $y'' + (e^{-2x} - \frac{1}{9})y = 0 \quad (e^{-x} = z)$

5. Two-parameter ODE
$x^2y'' + xy' + (\lambda^2x^2 - \nu^2)y = 0 \quad (\lambda x = z)$

6. $x^2y'' + \frac{1}{4}(x + \frac{3}{4})\,y = 0 \quad (y = u\sqrt{x},\ \sqrt{x} = z)$

7. $x^2y'' + xy' + \frac{1}{4}(x^2 - 1)y = 0 \quad (x = 2z)$

8. $(2x + 1)^2y'' + 2(2x + 1)y' + 16x(x + 1)y = 0$
$(2x + 1 = z)$

9. $xy'' + (2\nu + 1)y' + xy = 0 \quad (y = x^{-\nu}u)$

10. $x^2y'' + (1 - 2\nu)xy' + \nu^2(x^{2\nu} + 1 - \nu^2)y = 0$
$(y = x^\nu u,\ x^\nu = z)$

11. CAS EXPERIMENT. Change of Coefficient. Find and graph (on common axes) the solutions of

$$y'' + kx^{-1}y' + y = 0,\ y(0) = 1,\ y'(0) = 0,$$

for $k = 0, 1, 2, \cdots, 10$ (or as far as you get useful graphs). For what k do you get elementary functions? Why? Try for noninteger k, particularly between 0 and 2, to see the continuous change of the curve. Describe the change of the location of the zeros and of the extrema as k increases from 0. Can you interpret the ODE as a model in mechanics, thereby explaining your observations?

12. CAS EXPERIMENT. Bessel Functions for Large x.

(a) Graph $J_n(x)$ for $n = 0, \cdots, 5$ on common axes.

(b) Experiment with (14) for integer n. Using graphs, find out from which $x = x_n$ on the curves of (11) and (14) practically coincide. How does x_n change with n?

(c) What happens in (b) if $n = \pm\frac{1}{2}$? (Our usual notation in this case would be ν.)

(d) How does the error of (14) behave as a function of x for fixed n? [Error = exact value minus approximation (14).]

(e) Show from the graphs that $J_0(x)$ has extrema where $J_1(x) = 0$. Which formula proves this? Find further relations between zeros and extrema.

13–15 ZEROS of Bessel functions play a key role in modeling (e.g. of vibrations; see Sec. 12.9).

13. Interlacing of zeros. Using (21) and Rolle's theorem, show that between any two consecutive positive zeros of $J_n(x)$ there is precisely one zero of $J_{n+1}(x)$.

14. Zeros. Compute the first four positive zeros of $J_0(x)$ and $J_1(x)$ from (14). Determine the error and comment.

15. Interlacing of zeros. Using (21) and Rolle's theorem, show that between any two consecutive zeros of $J_0(x)$ there is precisely one zero of $J_1(x)$.

16–18 HALF-INTEGER PARAMETER: APPROACH BY THE ODE

16. Elimination of first derivative. Show that $y = uv$ with $v(x) = \exp\left(-\frac{1}{2}\int p(x)\,dx\right)$ gives from the ODE $y'' + p(x)y' + q(x)y = 0$ the ODE

$$u'' + \left[q(x) - \tfrac{1}{4}p(x)^2 - \tfrac{1}{2}p'(x)\right]u = 0,$$

not containing the first derivative of u.

17. Bessel's equation. Show that for (1) the substitution in Prob. 16 is $y = ux^{-1/2}$ and gives

(27) $$x^2u'' + (x^2 + \tfrac{1}{4} - \nu^2)u = 0.$$

18. Elementary Bessel functions. Derive (22) in Example 3 from (27).

19–25 APPLICATION OF (21): DERIVATIVES, INTEGRALS

Use the powerful formulas (21) to do Probs. 19–25. Show the details of your work.

19. Derivatives. Show that $J_0'(x) = -J_1(x)$, $J_1'(x) = J_0(x) - J_1(x)/x$, $J_2'(x) = \tfrac{1}{2}[J_1(x) - J_3(x)]$.

20. Bessel's equation. Derive (1) from (21).

21. Basic integral formula. Show that

$$\int x^\nu J_{\nu-1}(x)\,dx = x^\nu J_\nu(x) + c.$$

22. Basic integral formulas. Show that

$$\int x^{-\nu} J_{\nu+1}(x)\,dx = -x^{-\nu} J_\nu(x) + c,$$

$$\int J_{\nu+1}(x)\,dx = \int J_{\nu-1}(x)\,dx - 2J_\nu(x).$$

23. Integration. Show that $\int x^2 J_0(x)\,dx = x^2 J_1(x) + xJ_0(x) - \int J_0(x)\,dx$. (The last integral is nonelementary; tables exist, e.g., in Ref. [A13] in App. 1.)

24. Integration. Evaluate $\int x^{-1} J_4(x)\,dx$.

25. Integration. Evaluate $\int J_5(x)\,dx$.

5.5 Bessel Functions $Y_\nu(x)$. General Solution

To obtain a general solution of Bessel's equation (1), Sec. 5.4, for any ν, we now introduce **Bessel functions of the second kind** $Y_\nu(x)$, beginning with the case $\nu = n = 0$.

When $n = 0$, Bessel's equation can be written (divide by x)

(1) $$xy'' + y' + xy = 0.$$

Then the indicial equation (4) in Sec. 5.4 has a double root $r = 0$. This is Case 2 in Sec. 5.3. In this case we first have only one solution, $J_0(x)$. From (8) in Sec. 5.3 we see that the desired second solution must be of the form

(2) $$y_2(x) = J_0(x)\ln x + \sum_{m=1}^{\infty} A_m x^m.$$

We substitute y_2 and its derivatives

$$y_2' = J_0' \ln x + \frac{J_0}{x} + \sum_{m=1}^{\infty} mA_m x^{m-1}$$

$$y_2'' = J_0'' \ln x + \frac{2J_0'}{x} - \frac{J_0}{x^2} + \sum_{m=1}^{\infty} m(m-1)A_m x^{m-2}$$

into (1). Then the sum of the three logarithmic terms $xJ_0'' \ln x$, $J_0' \ln x$, and $xJ_0 \ln x$ is zero because J_0 is a solution of (1). The terms $-J_0/x$ and J_0/x (from xy'' and y') cancel. Hence we are left with

$$2J_0' + \sum_{m=1}^{\infty} m(m-1)A_m x^{m-1} + \sum_{m=1}^{\infty} mA_m x^{m-1} + \sum_{m=1}^{\infty} A_m x^{m+1} = 0.$$

Addition of the first and second series gives $\Sigma m^2 A_m x^{m-1}$. The power series of $J_0'(x)$ is obtained from (12) in Sec. 5.4 and the use of $m!/m = (m - 1)!$ in the form

$$J_0'(x) = \sum_{m=1}^{\infty} \frac{(-1)^m 2m x^{2m-1}}{2^{2m} (m!)^2} = \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m-1} m! \, (m-1)!}.$$

Together with $\Sigma m^2 A_m x^{m-1}$ and $\Sigma A_m x^{m+1}$ this gives

$$\text{(3*)} \qquad \sum_{m=1}^{\infty} \frac{(-1)^m x^{2m-1}}{2^{2m-2} m! \, (m-1)!} + \sum_{m=1}^{\infty} m^2 A_m x^{m-1} + \sum_{m=1}^{\infty} A_m x^{m+1} = 0.$$

First, we show that the A_m with odd subscripts are all zero. The power x^0 occurs only in the second series, with coefficient A_1. Hence $A_1 = 0$. Next, we consider the even powers x^{2s}. The first series contains none. In the second series, $m - 1 = 2s$ gives the term $(2s + 1)^2 A_{2s+1} x^{2s}$. In the third series, $m + 1 = 2s$. Hence by equating the sum of the coefficients of x^{2s} to zero we have

$$(2s + 1)^2 A_{2s+1} + A_{2s-1} = 0, \qquad s = 1, 2, \cdots.$$

Since $A_1 = 0$, we thus obtain $A_3 = 0, A_5 = 0, \cdots$, successively.

We now equate the sum of the coefficients of x^{2s+1} to zero. For $s = 0$ this gives

$$-1 + 4A_2 = 0, \qquad \text{thus} \qquad A_2 = \tfrac{1}{4}.$$

For the other values of s we have in the first series in (3*) $2m - 1 = 2s + 1$, hence $m = s + 1$, in the second $m - 1 = 2s + 1$, and in the third $m + 1 = 2s + 1$. We thus obtain

$$\frac{(-1)^{s+1}}{2^{2s}(s + 1)! \, s!} + (2s + 2)^2 A_{2s+2} + A_{2s} = 0.$$

For $s = 1$ this yields

$$\tfrac{1}{8} + 16A_4 + A_2 = 0, \qquad \text{thus} \qquad A_4 = -\tfrac{3}{128}$$

and in general

$$\text{(3)} \qquad A_{2m} = \frac{(-1)^{m-1}}{2^{2m}(m!)^2} \left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{m}\right), \qquad m = 1, 2, \cdots.$$

Using the short notations

$$\text{(4)} \qquad h_1 = 1 \qquad h_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m} \qquad m = 2, 3, \cdots$$

and inserting (4) and $A_1 = A_3 = \cdots = 0$ into (2), we obtain the result

$$y_2(x) = J_0(x) \ln x + \sum_{m=1}^{\infty} \frac{(-1)^{m-1} h_m}{2^{2m}(m!)^2} x^{2m}$$

$$\text{(5)} \qquad = J_0(x) \ln x + \frac{1}{4} x^2 - \frac{3}{128} x^4 + \frac{11}{13{,}824} x^6 - + \cdots.$$

Since J_0 and y_2 are linearly independent functions, they form a basis of (1) for $x > 0$. Of course, another basis is obtained if we replace y_2 by an independent particular solution of the form $a(y_2 + bJ_0)$, where $a\ (\neq 0)$ and b are constants. It is customary to choose $a = 2/\pi$ and $b = \gamma - \ln 2$, where the number $\gamma = 0.57721566490 \cdots$ is the so-called **Euler constant**, which is defined as the limit of

$$1 + \frac{1}{2} + \cdots + \frac{1}{s} - \ln s$$

as s approaches infinity. The standard particular solution thus obtained is called the **Bessel function of the second kind** *of order zero* (Fig. 112) or **Neumann's function** *of order zero* and is denoted by $Y_0(x)$. Thus [see (4)]

$$\text{(6)} \qquad Y_0(x) = \frac{2}{\pi}\left[J_0(x)\left(\ln\frac{x}{2} + \gamma\right) + \sum_{m=1}^{\infty} \frac{(-1)^{m-1}h_m}{2^{2m}(m!)^2} x^{2m}\right].$$

For small $x > 0$ the function $Y_0(x)$ behaves about like $\ln x$ (see Fig. 112, why?), and $Y_0(x) \to -\infty$ as $x \to 0$.

Bessel Functions of the Second Kind $Y_n(x)$

For $\nu = n = 1, 2, \cdots$ a second solution can be obtained by manipulations similar to those for $n = 0$, starting from (10), Sec. 5.4. It turns out that in these cases the solution also contains a logarithmic term.

The situation is not yet completely satisfactory, because the second solution is defined differently, depending on whether the order ν is an integer or not. To provide uniformity of formalism, it is desirable to adopt a form of the second solution that is valid for all values of the order. For this reason we introduce a standard second solution $Y_\nu(x)$ defined for all ν by the formula

$$\text{(7)} \qquad \begin{aligned} &\text{(a)} \quad Y_\nu(x) = \frac{1}{\sin \nu\pi}[J_\nu(x) \cos \nu\pi - J_{-\nu}(x)] \\ &\text{(b)} \quad Y_n(x) = \lim_{\nu \to n} Y_\nu(x). \end{aligned}$$

This function is called the **Bessel function of the second kind** *of order* ν or **Neumann's function**[7] *of order* ν. Figure 112 shows $Y_0(x)$ and $Y_1(x)$.

Let us show that J_ν and Y_ν are indeed linearly independent for all ν (and $x > 0$).

For noninteger order ν, the function $Y_\nu(x)$ is evidently a solution of Bessel's equation because $J_\nu(x)$ and $J_{-\nu}(x)$ are solutions of that equation. Since for those ν the solutions J_ν and $J_{-\nu}$ are linearly independent and Y_ν involves $J_{-\nu}$, the functions J_ν and Y_ν are

---

[7] CARL NEUMANN (1832–1925), German mathematician and physicist. His work on potential theory using integer equation methods inspired VITO VOLTERRA (1800–1940) of Rome, ERIK IVAR FREDHOLM (1866–1927) of Stockholm, and DAVID HILBERT (1962–1943) of Göttingen (see the footnote in Sec. 7.9) to develop the field of integral equations. For details see Birkhoff, G. and E. Kreyszig, The Establishment of Functional Analysis, *Historia Mathematica* 11 (1984), pp. 258–321.

The solutions $Y_\nu(x)$ are sometimes denoted by $N_\nu(x)$; in Ref. [A13] they are called **Weber's functions**; Euler's constant in (6) is often denoted by C or $\ln \gamma$.

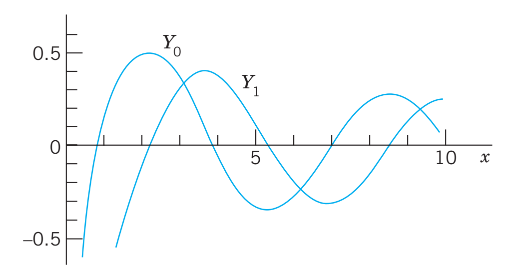


Fig. 112. Bessel functions of the second kind Y_0 and Y_1. (For a small table, see App. 5.)

linearly independent. Furthermore, it can be shown that the limit in (7b) exists and Y_n is a solution of Bessel's equation for integer order; see Ref. [A13] in App. 1. We shall see that the series development of $Y_n(x)$ contains a logarithmic term. Hence $J_n(x)$ and $Y_n(x)$ are linearly independent solutions of Bessel's equation. The series development of $Y_n(x)$ can be obtained if we insert the series (20) in Sec. 5.4 and (2) in this section for $J_\nu(x)$ and $J_{-\nu}(x)$ into (7a) and then let ν approach n; for details see Ref. [A13]. The result is

(8)
$$Y_n(x) = \frac{2}{\pi} J_n(x)\left(\ln\frac{x}{2} + \gamma\right) + \frac{x^n}{\pi}\sum_{m=0}^{\infty} \frac{(-1)^{m-1}(h_m + h_{m+n})}{2^{2m+n} m!\,(m+n)!} x^{2m}$$
$$- \frac{x^{-n}}{\pi}\sum_{m=0}^{n-1} \frac{(n-m-1)!}{2^{2m-n} m!} x^{2m}$$

where $x > 0$, $n = 0, 1, \cdots$, and [as in (4)] $h_0 = 0$, $h_1 = 1$,

$$h_m = 1 + \frac{1}{2} + \cdots + \frac{1}{m}, \qquad h_{m+n} = 1 + \frac{1}{2} + \cdots + \frac{1}{m+n}.$$

For $n = 0$ the last sum in (8) is to be replaced by 0 [giving agreement with (6)].

Furthermore, it can be shown that

$$Y_{-n}(x) = (-1)^n Y_n(x).$$

Our main result may now be formulated as follows.

THEOREM 1

General Solution of Bessel's Equation

A general solution of Bessel's equation for all values of ν (and $x > 0$) is

(9)
$$y(x) = C_1 J_\nu(x) + C_2 Y_\nu(x).$$

We finally mention that there is a practical need for solutions of Bessel's equation that are complex for real values of x. For this purpose the solutions

(10)
$$H_\nu^{(1)}(x) = J_\nu(x) + iY_\nu(x)$$
$$H_\nu^{(2)}(x) = J_\nu(x) - iY_\nu(x)$$

are frequently used. These linearly independent functions are called **Bessel functions of the third kind** *of order* ν or *first and second* **Hankel functions**[8] *of order* ν.

This finishes our discussion on Bessel functions, except for their "orthogonality," which we explain in Sec. 11.6. Applications to vibrations follow in Sec. 12.10.

PROBLEM SET 5.5

1–9 FURTHER ODE's REDUCIBLE TO BESSEL'S ODE

Find a general solution in terms of J_ν and Y_ν. Indicate whether you could also use $J_{-\nu}$ instead of Y_ν. Use the indicated substitution. Show the details of your work.

1. $x^2y'' + xy' + (x^2 - 16)\, y = 0$

2. $xy'' + 5y' + xy = 0 \quad (y = u/x^2)$

3. $9x^2y'' + 9xy' + (36x^4 - 16)y = 0 \quad (x^2 = z)$

4. $y'' + xy = 0 \quad (y = u\sqrt{x}, \frac{2}{3}x^{3/2} = z)$

5. $4xy'' + 4y' + y = 0 \quad (\sqrt{x} = z)$

6. $xy'' + y' + 36y = 0 \quad (12\sqrt{x} = z)$

7. $y'' + k^2x^2y = 0 \quad (y = u\sqrt{x}, \frac{1}{2}kx^2 = z)$

8. $y'' + k^2x^4y = 0 \quad (y = u\sqrt{x}, \frac{1}{3}kx^3 = z)$

9. $xy'' - 5y' + xy = 0 \quad (y = x^3u)$

10. CAS EXPERIMENT. Bessel Functions for Large x. It can be shown that for large x,

$$Y_n(x) \sim \sqrt{2/(\pi x)}\, \sin\left(x - \tfrac{1}{2}n\pi - \tfrac{1}{4}\pi\right) \tag{11}$$

with $\sim$ defined as in (14) of Sec. 5.4.

(a) Graph $Y_n(x)$ for $n = 0, \cdots, 5$ on common axes. Are there relations between zeros of one function and extrema of another? For what functions?

(b) Find out from graphs from which $x = x_n$ on the curves of (8) and (11) (both obtained from your CAS) practically coincide. How does x_n change with n?

(c) Calculate the first ten zeros x_m, $m = 1, \cdots, 10$, of $Y_0(x)$ from your CAS and from (11). How does the error behave as m increases?

(d) Do (c) for $Y_1(x)$ and $Y_2(x)$. How do the errors compare to those in (c)?

11–15 HANKEL AND MODIFIED BESSEL FUNCTIONS

11. Hankel functions. Show that the Hankel functions (10) form a basis of solutions of Bessel's equation for any ν.

12. Modified Bessel functions of the first kind of order ν are defined by $I_\nu(x) = i^{-\nu}J_\nu(ix)$, $i = \sqrt{-1}$. Show that I_ν satisfies the ODE

$$x^2y'' + xy' - (x^2 + \nu^2)\, y = 0. \tag{12}$$

13. Modified Bessel functions. Show that $I_\nu(x)$ has the representation

$$I_\nu(x) = \sum_{m=0}^{\infty} \frac{x^{2m+\nu}}{2^{2m+\nu}m!\, \Gamma(m + \nu + 1)}. \tag{13}$$

14. Reality of I_ν. Show that $I_\nu(x)$ is real for all real x (and real ν), $I_\nu(x) \neq 0$ for all real $x \neq 0$, and $I_{-n}(x) = I_n(x)$, where n is any integer.

15. Modified Bessel functions of the third kind (sometimes called *of the second kind*) are defined by the formula (14) below. Show that they satisfy the ODE (12).

$$K_\nu(x) = \frac{\pi}{2 \sin \nu\pi}\left[I_{-\nu}(x) - I_\nu(x)\right]. \tag{14}$$

CHAPTER 5 REVIEW QUESTIONS AND PROBLEMS

1. Why are we looking for power series solutions of ODEs?

2. What is the difference between the two methods in this chapter? Why do we need two methods?

3. What is the indicial equation? Why is it needed?

4. List the three cases of the Frobenius method, and give examples of your own.

5. Write down the most important ODEs in this chapter from memory.

6. Can a power series solution reduce to a polynomial? When? Why is this important?

7. What is the hypergeometric equation? Where does the name come from?

8. List some properties of the Legendre polynomials.

9. Why did we introduce two kinds of Bessel functions?

10. Can a Bessel function reduce to an elementary function? When?

[8]HERMANN HANKEL (1839–1873), German mathematician.

11–20 **POWER SERIES METHOD OR FROBENIUS METHOD**

Find a basis of solutions. Try to identify the series as expansions of known functions. Show the details of your work.

11. $y'' + 4y = 0$

12. $xy'' + (1 - 2x)y' + (x - 1)y = 0$

13. $(x - 1)^2 y'' - (x - 1)y' - 35y = 0$

14. $16(x + 1)^2 y'' + 3y = 0$

15. $x^2 y'' + xy' + (x^2 - 5)y = 0$

16. $x^2 y'' + 2x^3 y' + (x^2 - 2)y = 0$

17. $xy'' - (x + 1)y' + y = 0$

18. $xy'' + 3y' + 4x^3 y = 0$

19. $y'' + \dfrac{1}{4x} y = 0$

20. $xy'' + y' - xy = 0$

SUMMARY OF CHAPTER 5

Series Solution of ODEs. Special Functions

The **power series method** gives solutions of linear ODEs

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

with ***variable coefficients*** p and q in the form of a power series (with any center x_0, e.g., $x_0 = 0$)

$$y(x) = \sum_{m=0}^{\infty} a_m (x - x_0)^m = a_0 + a_1(x - x_0) + a_2(x - x_0)^2 + \cdots. \tag{2}$$

Such a solution is obtained by substituting (2) and its derivatives into (1). This gives a ***recurrence formula*** for the coefficients. You may program this formula (or even obtain and graph the whole solution) on your CAS.

If p and q are **analytic** at x_0 (that is, representable by a power series in powers of $x - x_0$ with positive radius of convergence; Sec. 5.1), then (1) has solutions of this form (2). The same holds if $\tilde{h}$, $\tilde{p}$, $\tilde{q}$ in

$$\tilde{h}(x)y'' + \tilde{p}(x)y' + \tilde{q}(x)y = 0$$

are analytic at x_0 and $\tilde{h}(x_0) \neq 0$, so that we can divide by $\tilde{h}$ and obtain the standard form (1). ***Legendre's equation*** is solved by the power series method in Sec. 5.2.

The ***Frobenius method*** (Sec. 5.3) extends the power series method to ODEs

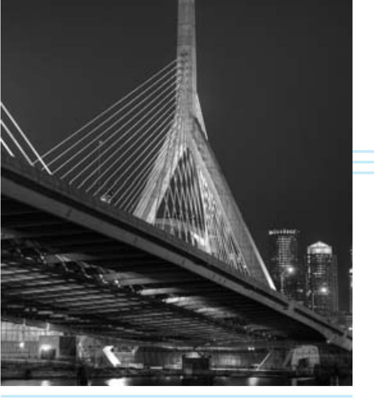
$$y'' + \frac{a(x)}{x - x_0} y' + \frac{b(x)}{(x - x_0)^2} y = 0 \tag{3}$$

whose coefficients are **singular** (i.e., not analytic) at x_0, but are "not too bad," namely, such that a and b are analytic at x_0. Then (3) has at least one solution of the form

$$y(x) = (x - x_0)^r \sum_{m=0}^{\infty} a_m (x - x_0)^m = a_0(x - x_0)^r + a_1(x - x_0)^{r+1} + \cdots \tag{4}$$

where r can be any real (or even complex) number and is determined by substituting (4) into (3) from the ***indicial equation*** (Sec. 5.3), along with the coefficients of (4). A second linearly independent solution of (3) may be of a similar form (with different r and a_m's) or may involve a logarithmic term. ***Bessel's equation*** is solved by the Frobenius method in Secs. 5.4 and 5.5.

"**Special functions**" is a common name for higher functions, as opposed to the usual functions of calculus. Most of them arise either as nonelementary integrals [see (24)–(44) in App. 3.1] or as solutions of (1) or (3). They get a name and notation and are included in the usual CASs if they are important in application or in theory. Of this kind, and particularly useful to the engineer and physicist, are ***Legendre's equation and polynomials*** $P_0, P_1, \cdots$ (Sec. 5.2), ***Gauss's hypergeometric equation and functions*** $F(a, b, c; x)$ (Sec. 5.3), and ***Bessel's equation and functions*** J_ν ***and*** Y_ν (Secs. 5.4, 5.5).



CHAPTER 6

Laplace Transforms

Laplace transforms are invaluable for any engineer's mathematical toolbox as they make solving linear ODEs and related initial value problems, as well as systems of linear ODEs, much easier. Applications abound: electrical networks, springs, mixing problems, signal processing, and other areas of engineering and physics.

The process of solving an ODE using the Laplace transform method consists of three steps, shown schematically in Fig. 113:

Step 1. The given ODE is transformed into an algebraic equation, called the **subsidiary equation**.

Step 2. The subsidiary equation is solved by purely algebraic manipulations.

Step 3. The solution in Step 2 is transformed back, resulting in the solution of the given problem.

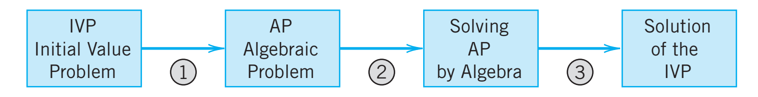


Fig. 113. Solving an IVP by Laplace transforms

The key motivation for learning about Laplace transforms is that the process of solving an ODE is simplified to an algebraic problem (and transformations). This type of mathematics that converts problems of calculus to algebraic problems is known as *operational calculus*. The Laplace transform method has two main advantages over the methods discussed in Chaps. 1–4:

I. Problems are solved more directly: Initial value problems are solved without first determining a general solution. Nonhomogenous ODEs are solved without first solving the corresponding homogeneous ODE.

II. More importantly, the use of the ***unit step function*** (***Heaviside function*** in Sec. 6.3) and ***Dirac's delta*** (in Sec. 6.4) make the method particularly powerful for problems with inputs (driving forces) that have discontinuities or represent short impulses or complicated periodic functions.

The following chart shows where to find information on the Laplace transform in this book.

Topic	Where to find it
ODEs, engineering applications and Laplace transforms	Chapter 6
PDEs, engineering applications and Laplace transforms	Section 12.11
List of general formulas of Laplace transforms	Section 6.8
List of Laplace transforms and inverses	Section 6.9
Note: Your CAS can handle most Laplace transforms.	

Prerequisite: Chap. 2
Sections that may be omitted in a shorter course: 6.5, 6.7
References and Answers to Problems: App. 1 Part A, App. 2.

6.1 Laplace Transform. Linearity. First Shifting Theorem (*s*-Shifting)

In this section, we learn about Laplace transforms and some of their properties. Because Laplace transforms are of basic importance to the engineer, the student should pay close attention to the material. Applications to ODEs follow in the next section.

Roughly speaking, the Laplace transform, when applied to a function, changes that function into a new function by using a process that involves integration. Details are as follows.

If $f(t)$ is a function defined for all $t \geqq 0$, its **Laplace transform**[1] is the integral of $f(t)$ times e^{-st} from $t = 0$ to ∞. It is a function of s, say, $F(s)$, and is denoted by $\mathcal{L}(f)$; thus

(1) $$F(s) = \mathcal{L}(f) = \int_0^\infty e^{-st} f(t)\, dt.$$

Here we must assume that $f(t)$ is such that the integral exists (that is, has some finite value). This assumption is usually satisfied in applications—we shall discuss this near the end of the section.

---

[1] PIERRE SIMON MARQUIS DE LAPLACE (1749–1827), great French mathematician, was a professor in Paris. He developed the foundation of potential theory and made important contributions to celestial mechanics, astronomy in general, special functions, and probability theory. Napoléon Bonaparte was his student for a year. For Laplace's interesting political involvements, see Ref. [GenRef2], listed in App. 1.

The powerful practical Laplace transform techniques were developed over a century later by the English electrical engineer OLIVER HEAVISIDE (1850–1925) and were often called "Heaviside calculus."

We shall drop variables when this simplifies formulas without causing confusion. For instance, in (1) we wrote $\mathcal{L}(f)$ instead of $\mathcal{L}(f)(s)$ and in (1*) $\mathcal{L}^{-1}(F)$ instead of $\mathcal{L}^{-1}(F)(t)$.

Not only is the result $F(s)$ called the Laplace transform, but the operation just described, which yields $F(s)$ from a given $f(t)$, is also called the **Laplace transform**. It is an "**integral transform**"

$$F(s) = \int_0^{\infty} k(s, t) f(t)\, dt$$

with "**kernel**" $k(s, t) = e^{-st}$.

Note that the Laplace transform is called an integral transform because it transforms (changes) a function in one space to a function in another space by a *process of integration* that involves a kernel. The kernel or kernel function is a function of the variables in the two spaces and defines the integral transform.

Furthermore, the given function $f(t)$ in (1) is called the **inverse transform** of $F(s)$ and is denoted by $\mathcal{L}^{-1}(F)$; that is, we shall write

$$f(t) = \mathcal{L}^{-1}(F). \tag{1*}$$

Note that (1) and (1*) together imply $\mathcal{L}^{-1}(\mathcal{L}(f)) = f$ and $\mathcal{L}(\mathcal{L}^{-1}(F)) = F$.

Notation

Original functions depend on t and their transforms on s—keep this in mind! Original functions are denoted by ***lowercase letters*** and their transforms by the same ***letters in capital***, so that $F(s)$ denotes the transform of $f(t)$, and $Y(s)$ denotes the transform of $y(t)$, and so on.

EXAMPLE 1 Laplace Transform

Let $f(t) = 1$ when $t \geqq 0$. Find $F(s)$.

Solution. From (1) we obtain by integration

$$\mathcal{L}(f) = \mathcal{L}(1) = \int_0^{\infty} e^{-st}\, dt = -\frac{1}{s} e^{-st} \Big|_0^{\infty} = \frac{1}{s} \qquad (s > 0).$$

Such an integral is called an **improper integral** and, by definition, is evaluated according to the rule

$$\int_0^{\infty} e^{-st} f(t)\, dt = \lim_{T \to \infty} \int_0^{T} e^{-st} f(t)\, dt.$$

Hence our convenient notation means

$$\int_0^{\infty} e^{-st}\, dt = \lim_{T \to \infty} \left[-\frac{1}{s} e^{-st} \right]_0^{T} = \lim_{T \to \infty} \left[-\frac{1}{s} e^{-sT} + \frac{1}{s} e^{0} \right] = \frac{1}{s} \qquad (s > 0).$$

We shall use this notation throughout this chapter. ■

EXAMPLE 2 Laplace Transform $\mathcal{L}(e^{at})$ of the Exponential Function e^{at}

Let $f(t) = e^{at}$ when $t \geqq 0$, where a is a constant. Find $\mathcal{L}(f)$.

Solution. Again by (1),

$$\mathcal{L}(e^{at}) = \int_0^{\infty} e^{-st} e^{at}\, dt = \frac{1}{a - s} e^{-(s-a)t} \Big|_0^{\infty};$$

hence, when $s - a > 0$,

$$\mathcal{L}(e^{at}) = \frac{1}{s - a}. \qquad ■$$

Must we go on in this fashion and obtain the transform of one function after another directly from the definition? No! We can obtain new transforms from known ones by the use of the many general properties of the Laplace transform. Above all, the Laplace transform is a "linear operation," just as are differentiation and integration. By this we mean the following.

THEOREM 1

Linearity of the Laplace Transform

The Laplace transform is a linear operation; that is, for any functions $f(t)$ and $g(t)$ whose transforms exist and any constants a and b the transform of $af(t) + bg(t)$ exists, and

$$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}.$$

PROOF This is true because integration is a linear operation so that (1) gives

$$\begin{aligned}\mathcal{L}\{af(t) + bg(t)\} &= \int_0^\infty e^{-st}[af(t) + bg(t)]\,dt \\ &= a\int_0^\infty e^{-st}f(t)\,dt + b\int_0^\infty e^{-st}g(t)\,dt = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}. \quad \blacksquare\end{aligned}$$

EXAMPLE 3 **Application of Theorem 1: Hyperbolic Functions**

Find the transforms of cosh at and sinh at.

Solution. Since $\cosh at = \frac{1}{2}(e^{at} + e^{-at})$ and $\sinh at = \frac{1}{2}(e^{at} - e^{-at})$, we obtain from Example 2 and Theorem 1

$$\mathcal{L}(\cosh at) = \frac{1}{2}(\mathcal{L}(e^{at}) + \mathcal{L}(e^{-at})) = \frac{1}{2}\left(\frac{1}{s-a} + \frac{1}{s+a}\right) = \frac{s}{s^2 - a^2}$$

$$\mathcal{L}(\sinh at) = \frac{1}{2}(\mathcal{L}(e^{at}) - \mathcal{L}(e^{-at})) = \frac{1}{2}\left(\frac{1}{s-a} - \frac{1}{s+a}\right) = \frac{a}{s^2 - a^2}. \quad \blacksquare$$

EXAMPLE 4 **Cosine and Sine**

Derive the formulas

$$\mathcal{L}(\cos \omega t) = \frac{s}{s^2 + \omega^2}, \qquad \mathcal{L}(\sin \omega t) = \frac{\omega}{s^2 + \omega^2}.$$

Solution. We write $L_c = \mathcal{L}(\cos \omega t)$ and $L_s = \mathcal{L}(\sin \omega t)$. Integrating by parts and noting that the integral-free parts give no contribution from the upper limit ∞, we obtain

$$L_c = \int_0^\infty e^{-st}\cos \omega t\,dt = \frac{e^{-st}}{-s}\cos \omega t\Big|_0^\infty - \frac{\omega}{s}\int_0^\infty e^{-st}\sin \omega t\,dt = \frac{1}{s} - \frac{\omega}{s}L_s,$$

$$L_s = \int_0^\infty e^{-st}\sin \omega t\,dt = \frac{e^{-st}}{-s}\sin \omega t\Big|_0^\infty + \frac{\omega}{s}\int_0^\infty e^{-st}\cos \omega t\,dt = \frac{\omega}{s}L_c.$$

By substituting L_s into the formula for L_c on the right and then by substituting L_c into the formula for L_s on the right, we obtain

$$L_c = \frac{1}{s} - \frac{\omega}{s}\left(\frac{\omega}{s}L_c\right), \qquad L_c\left(1 + \frac{\omega^2}{s^2}\right) = \frac{1}{s}, \qquad L_c = \frac{s}{s^2 + \omega^2},$$

$$L_s = \frac{\omega}{s}\left(\frac{1}{s} - \frac{\omega}{s}L_s\right), \qquad L_s\left(1 + \frac{\omega^2}{s^2}\right) = \frac{\omega}{s^2}, \qquad L_s = \frac{\omega}{s^2 + \omega^2}.$$

Basic transforms are listed in Table 6.1. We shall see that from these almost all the others can be obtained by the use of the general properties of the Laplace transform. Formulas 1–3 are special cases of formula 4, which is proved by induction. Indeed, it is true for $n = 0$ because of Example 1 and $0! = 1$. We make the induction hypothesis that it holds for any integer $n \geqq 0$ and then get it for $n + 1$ directly from (1). Indeed, integration by parts first gives

$$\mathcal{L}(t^{n+1}) = \int_0^\infty e^{-st} t^{n+1}\, dt = -\frac{1}{s} e^{-st} t^{n+1} \Big|_0^\infty + \frac{n+1}{s} \int_0^\infty e^{-st} t^n \, dt.$$

Now the integral-free part is zero and the last part is $(n + 1)/s$ times $\mathcal{L}(t^n)$. From this and the induction hypothesis,

$$\mathcal{L}(t^{n+1}) = \frac{n+1}{s}\mathcal{L}(t^n) = \frac{n+1}{s} \cdot \frac{n!}{s^{n+1}} = \frac{(n+1)!}{s^{n+2}}.$$

This proves formula 4.

Table 6.1 Some Functions $f(t)$ and Their Laplace Transforms $\mathcal{L}(f)$

	$f(t)$	$\mathcal{L}(f)$		$f(t)$	$\mathcal{L}(f)$
1	1	$1/s$	7	$\cos \omega t$	$\frac{s}{s^2 + \omega^2}$
2	t	$1/s^2$	8	$\sin \omega t$	$\frac{\omega}{s^2 + \omega^2}$
3	t^2	$2!/s^3$	9	$\cosh at$	$\frac{s}{s^2 - a^2}$
4	t^n $(n = 0, 1, \cdots)$	$\frac{n!}{s^{n+1}}$	10	$\sinh at$	$\frac{a}{s^2 - a^2}$
5	t^a (a positive)	$\frac{\Gamma(a+1)}{s^{a+1}}$	11	$e^{at} \cos \omega t$	$\frac{s-a}{(s-a)^2 + \omega^2}$
6	e^{at}	$\frac{1}{s-a}$	12	$e^{at} \sin \omega t$	$\frac{\omega}{(s-a)^2 + \omega^2}$

$\Gamma(a+1)$ in formula 5 is the so-called *gamma function* [(15) in Sec. 5.5 or (24) in App. A3.1]. We get formula 5 from (1), setting $st = x$:

$$\mathcal{L}(t^a) = \int_0^\infty e^{-st} t^a \, dt = \int_0^\infty e^{-x} \left(\frac{x}{s}\right)^a \frac{dx}{s} = \frac{1}{s^{a+1}} \int_0^\infty e^{-x} x^a \, dx$$

where $s > 0$. The last integral is precisely that defining $\Gamma(a+1)$, so we have $\Gamma(a+1)/s^{a+1}$, as claimed. (**CAUTION!** $\Gamma(a+1)$ has x^a in the integral, not x^{a+1}.)

Note the formula 4 also follows from 5 because $\Gamma(n+1) = n!$ for integer $n \geqq 0$.

Formulas 6–10 were proved in Examples 2–4. Formulas 11 and 12 will follow from 7 and 8 by "shifting," to which we turn next.

s-Shifting: Replacing *s* by *s* − *a* in the Transform

The Laplace transform has the very useful property that, if we know the transform of $f(t)$, we can immediately get that of $e^{at} f(t)$, as follows.

THEOREM 2

First Shifting Theorem, *s*-Shifting

If $f(t)$ has the transform $F(s)$ (where $s > k$ for some k), then $e^{at} f(t)$ has the transform $F(s-a)$ (where $s - a > k$). In formulas,

$$\mathcal{L}\{e^{at} f(t)\} = F(s-a)$$

or, if we take the inverse on both sides,

$$e^{at} f(t) = \mathcal{L}^{-1}\{F(s-a)\}.$$

PROOF We obtain $F(s-a)$ by replacing s with $s - a$ in the integral in (1), so that

$$F(s-a) = \int_0^\infty e^{-(s-a)t} f(t) \, dt = \int_0^\infty e^{-st} [e^{at} f(t)] \, dt = \mathcal{L}\{e^{at} f(t)\}.$$

If $F(s)$ exists (i.e., is finite) for s greater than some k, then our first integral exists for $s - a > k$. Now take the inverse on both sides of this formula to obtain the second formula in the theorem. (**CAUTION!** $-a$ in $F(s-a)$ but $+a$ in $e^{at} f(t)$.) ■

EXAMPLE 5 **s-Shifting: Damped Vibrations. Completing the Square**

From Example 4 and the first shifting theorem we immediately obtain formulas 11 and 12 in Table 6.1,

$$\mathcal{L}\{e^{at} \cos \omega t\} = \frac{s-a}{(s-a)^2 + \omega^2}, \qquad \mathcal{L}\{e^{at} \sin \omega t\} = \frac{\omega}{(s-a)^2 + \omega^2}.$$

For instance, use these formulas to find the inverse of the transform

$$\mathcal{L}(f) = \frac{3s - 137}{s^2 + 2s + 401}.$$

Solution. Applying the inverse transform, using its linearity (Prob. 24), and completing the square, we obtain

$$f = \mathcal{L}^{-1}\left\{\frac{3(s+1)-140}{(s+1)^2+400}\right\} = 3\mathcal{L}^{-1}\left\{\frac{s+1}{(s+1)^2+20^2}\right\} - 7\mathcal{L}^{-1}\left\{\frac{20}{(s+1)^2+20^2}\right\}.$$

We now see that the inverse of the right side is the damped vibration (Fig. 114)

$$f(t) = e^{-t}(3\cos 20t - 7\sin 20t).$$

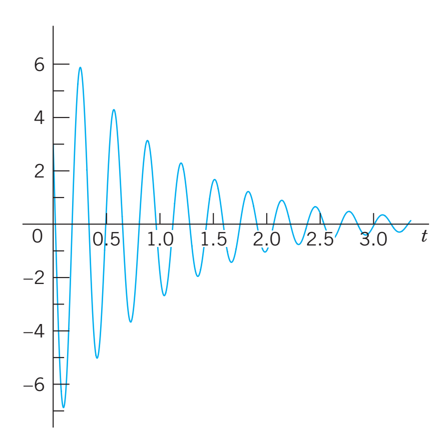


Fig. 114. Vibrations in Example 5

Existence and Uniqueness of Laplace Transforms

This is not a big *practical* problem because in most cases we can check the solution of an ODE without too much trouble. Nevertheless we should be aware of some basic facts.

A function $f(t)$ has a Laplace transform if it does not grow too fast, say, if for all $t \geqq 0$ and some constants M and k it satisfies the "**growth restriction**"

$$|f(t)| \leqq Me^{kt}. \tag{2}$$

(The growth restriction (2) is sometimes called "growth of exponential order," which may be misleading since it hides that the exponent must be kt, not kt^2 or similar.)

$f(t)$ need not be continuous, but it should not be too bad. The technical term (generally used in mathematics) is *piecewise continuity*. $f(t)$ is **piecewise continuous** on a finite interval $a \leqq t \leqq b$ where f is defined, if this interval can be divided into *finitely many* subintervals in each of which f is continuous and has finite limits as t approaches either endpoint of such a subinterval from the interior. This then gives **finite jumps** as in Fig. 115 as the only possible discontinuities, but this suffices in most applications, and so does the following theorem.

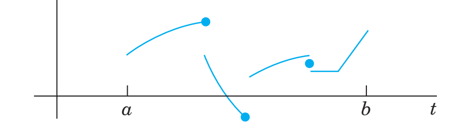


Fig. 115. Example of a piecewise continuous function $f(t)$. (The dots mark the function values at the jumps.)

THEOREM 3

Existence Theorem for Laplace Transforms

If $f(t)$ is defined and piecewise continuous on every finite interval on the semi-axis $t \geqq 0$ and satisfies (2) for all $t \geqq 0$ and some constants M and k, then the Laplace transform $\mathcal{L}(f)$ exists for all $s > k$.

PROOF Since $f(t)$ is piecewise continuous, $e^{-st}f(t)$ is integrable over any finite interval on the t-axis. From (2), assuming that $s > k$ (to be needed for the existence of the last of the following integrals), we obtain the proof of the existence of $\mathcal{L}(f)$ from

$$|\mathcal{L}(f)| = \left| \int_0^\infty e^{-st} f(t)\, dt \right| \leqq \int_0^\infty |f(t)| e^{-st}\, dt \leqq \int_0^\infty M e^{kt} e^{-st}\, dt = \frac{M}{s-k}. \qquad \blacksquare$$

Note that (2) can be readily checked. For instance, $\cosh t < e^t$, $t^n < n!e^t$ (because $t^n/n!$ is a single term of the Maclaurin series), and so on. A function that does not satisfy (2) for any M and k is e^{t^2} (take logarithms to see it). We mention that the conditions in Theorem 3 are sufficient rather than necessary (see Prob. 22).

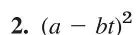
Uniqueness. If the Laplace transform of a given function exists, it is uniquely determined. Conversely, it can be shown that if two functions (both defined on the positive real axis) have the same transform, these functions cannot differ over an interval of positive length, although they may differ at isolated points (see Ref. [A14] in App. 1). Hence we may say that the inverse of a given transform is essentially unique. In particular, if two *continuous* functions have the same transform, they are completely identical.

PROBLEM SET 6.1

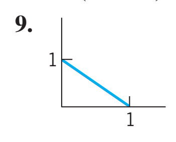
1–16 LAPLACE TRANSFORMS

Find the transform. Show the details of your work. Assume that a, b, ω, θ are constants.

1. $3t + 12$
2. $(a - bt)^2$

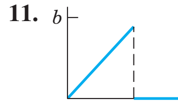


3. $\cos \pi t$
4. $\cos^2 \omega t$
5. $e^{2t} \sinh t$
6. $e^{-t} \sinh 4t$
7. $\sin(\omega t + \theta)$
8. $1.5 \sin(3t - \pi/2)$
9.

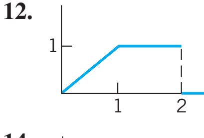


10.

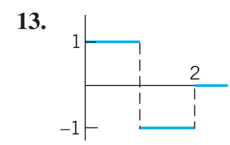
11.



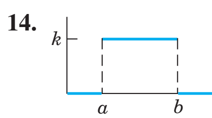
12.



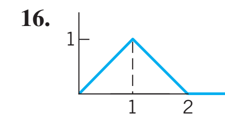
13.



14.



15.
16.



17–24 SOME THEORY

17. **Table 6.1.** Convert this table to a table for finding inverse transforms (with obvious changes, e.g., $\mathcal{L}^{-1}(1/s^n) = t^{n-1}/(n-1)$, etc).
18. Using $\mathcal{L}(f)$ in Prob. 10, find $\mathcal{L}(f_1)$, where $f_1(t) = 0$ if $t \leqq 2$ and $f_1(t) = 1$ if $t > 2$.
19. **Table 6.1.** Derive formula 6 from formulas 9 and 10.
20. **Nonexistence.** Show that e^{t^2} does not satisfy a condition of the form (2).
21. **Nonexistence.** Give simple examples of functions (defined for all $t \geqq 0$) that have no Laplace transform.
22. **Existence.** Show that $\mathcal{L}(1/\sqrt{t}) = \sqrt{\pi/s}$. [Use (30) $\Gamma(\frac{1}{2}) = \sqrt{\pi}$ in App. 3.1.] Conclude from this that the conditions in Theorem 3 are sufficient but not necessary for the existence of a Laplace transform.

23. Change of scale. If $\mathcal{L}(f(t)) = F(s)$ and c is any positive constant, show that $\mathcal{L}(f(ct)) = F(s/c)/c$ (***Hint:*** Use (1).) Use this to obtain $\mathcal{L}(\cos \omega t)$ from $\mathcal{L}(\cos t)$.

24. Inverse transform. Prove that $\mathcal{L}^{-1}$ is linear. *Hint:* Use the fact that $\mathcal{L}$ is linear.

25–32 INVERSE LAPLACE TRANSFORMS

Given $F(s) = \mathcal{L}(f)$, find $f(t)$. a, b, L, n are constants. Show the details of your work.

25. $\dfrac{0.2s + 1.8}{s^2 + 3.24}$

26. $\dfrac{5s + 1}{s^2 - 25}$

27. $\dfrac{s}{L^2 s^2 + n^2 \pi^2}$

28. $\dfrac{1}{(s + \sqrt{2})(s - \sqrt{3})}$

29. $\dfrac{12}{s^4} - \dfrac{228}{s^6}$

30. $\dfrac{4s + 32}{s^2 - 16}$

31. $\dfrac{s + 10}{s^2 - s - 2}$

32. $\dfrac{1}{(s + a)(s + b)}$

33–45 APPLICATION OF *s*-SHIFTING

In Probs. 33–36 find the transform. In Probs. 37–45 find the inverse transform. Show the details of your work.

33. $t^2 e^{-3t}$

34. $ke^{-at} \cos \omega t$

35. $0.5e^{-4.5t} \sin 2\pi t$

36. $\sinh t \cos t$

37. $\dfrac{\pi}{(s + \pi)^2}$

38. $\dfrac{6}{(s + 1)^3}$

39. $\dfrac{21}{(s + \sqrt{2})^4}$

40. $\dfrac{4}{s^2 - 2s - 3}$

41. $\dfrac{\pi}{s^2 + 10\pi s + 24\pi^2}$

42. $\dfrac{a_0}{s + 1} + \dfrac{a_1}{(s + 1)^2} + \dfrac{a_2}{(s + 1)^3}$

43. $\dfrac{2s - 1}{s^2 - 6s + 18}$

44. $\dfrac{a(s + k) + b\pi}{(s + k)^2 + \pi^2}$

45. $\dfrac{k_0(s + a) + k_1}{(s + a)^2}$

6.2 Transforms of Derivatives and Integrals. ODEs

The Laplace transform is a method of solving ODEs and initial value problems. The crucial idea is that ***operations of calculus on functions are replaced by operations of algebra on transforms***. Roughly, *differentiation* of $f(t)$ will correspond to *multiplication* of $\mathcal{L}(f)$ by s (see Theorems 1 and 2) and *integration* of $f(t)$ to *division* of $\mathcal{L}(f)$ by s. To solve ODEs, we must first consider the Laplace transform of derivatives. You have encountered such an idea in your study of logarithms. Under the application of the natural logarithm, a product of numbers becomes a sum of their logarithms, a division of numbers becomes their difference of logarithms (see Appendix 3, formulas (2), (3)). To simplify calculations was one of the main reasons that logarithms were invented in pre-computer times.

THEOREM 1

Laplace Transform of Derivatives

The transforms of the first and second derivatives of $f(t)$ satisfy

$$\mathcal{L}(f') = s\mathcal{L}(f) - f(0) \tag{1}$$

$$\mathcal{L}(f'') = s^2\mathcal{L}(f) - sf(0) - f'(0). \tag{2}$$

Formula (1) *holds if $f(t)$ is continuous for all $t \geqq 0$ and satisfies the growth restriction* (2) *in Sec.* 6.1 *and $f'(t)$ is piecewise continuous on every finite interval on the semi-axis $t \geqq 0$. Similarly,* (2) *holds if f and f' are continuous for all $t \geqq 0$ and satisfy the growth restriction and f'' is piecewise continuous on every finite interval on the semi-axis $t \geqq 0$.*

PROOF We prove (1) first under the *additional assumption* that f' is continuous. Then, by the definition and integration by parts,

$$\mathcal{L}(f') = \int_0^\infty e^{-st} f'(t)\,dt = \left[e^{-st} f(t)\right]\Big|_0^\infty + s\int_0^\infty e^{-st} f(t)\,dt.$$

Since f satisfies (2) in Sec. 6.1, the integrated part on the right is zero at the upper limit when $s > k$, and at the lower limit it contributes $-f(0)$. The last integral is $\mathcal{L}(f)$. It exists for $s > k$ because of Theorem 3 in Sec. 6.1. Hence $\mathcal{L}(f')$ exists when $s > k$ and (1) holds.

If f' is merely piecewise continuous, the proof is similar. In this case the interval of integration of f' must be broken up into parts such that f' is continuous in each such part.

The proof of (2) now follows by applying (1) to f'' and then substituting (1), that is

$$\mathcal{L}(f'') = s\mathcal{L}(f') - f'(0) = s[s\mathcal{L}(f) - f(0)] = s^2\mathcal{L}(f) - sf(0) - f'(0). \qquad \blacksquare$$

Continuing by substitution as in the proof of (2) and using induction, we obtain the following extension of Theorem 1.

THEOREM 2

Laplace Transform of the Derivative $f^{(n)}$ of Any Order

Let $f, f', \cdots, f^{(n-1)}$ be continuous for all $t \geqq 0$ and satisfy the growth restriction (2) in Sec. 6.1. Furthermore, let $f^{(n)}$ be piecewise continuous on every finite interval on the semi-axis $t \geqq 0$. Then the transform of $f^{(n)}$ satisfies

$$\textbf{(3)} \qquad \mathcal{L}(f^{(n)}) = s^n\mathcal{L}(f) - s^{n-1}f(0) - s^{n-2}f'(0) - \cdots - f^{(n-1)}(0).$$

EXAMPLE 1 **Transform of a Resonance Term (Sec. 2.8)**

Let $f(t) = t \sin \omega t$. Then $f(0) = 0$, $f'(t) = \sin \omega t + \omega t \cos \omega t$, $f'(0) = 0$, $f'' = 2\omega \cos \omega t - \omega^2 t \sin \omega t$. Hence by (2),

$$\mathcal{L}(f'') = 2\omega \frac{s}{s^2 + \omega^2} - \omega^2\mathcal{L}(f) = s^2\mathcal{L}(f), \qquad \text{thus} \qquad \mathcal{L}(f) = \mathcal{L}(t \sin \omega t) = \frac{2\omega s}{(s^2 + \omega^2)^2}. \qquad \blacksquare$$

EXAMPLE 2 **Formulas 7 and 8 in Table 6.1, Sec. 6.1**

This is a third derivation of $\mathcal{L}(\cos \omega t)$ and $\mathcal{L}(\sin \omega t)$; cf. Example 4 in Sec. 6.1. Let $f(t) = \cos \omega t$. Then $f(0) = 1$, $f'(0) = 0$, $f''(t) = -\omega^2 \cos \omega t$. From this and (2) we obtain

$$\mathcal{L}(f'') = s^2\mathcal{L}(f) - s = -\omega^2\mathcal{L}(f). \qquad \text{By algebra,} \qquad \mathcal{L}(\cos \omega t) = \frac{s}{s^2 + \omega^2}.$$

Similarly, let $g = \sin \omega t$. Then $g(0) = 0$, $g' = \omega \cos \omega t$. From this and (1) we obtain

$$\mathcal{L}(g') = s\mathcal{L}(g) = \omega\mathcal{L}(\cos \omega t). \qquad \text{Hence,} \qquad \mathcal{L}(\sin \omega t) = \frac{\omega}{s}\mathcal{L}(\cos \omega t) = \frac{\omega}{s^2 + \omega^2}. \qquad \blacksquare$$

Laplace Transform of the Integral of a Function

Differentiation and integration are inverse operations, and so are multiplication and division. Since differentiation of a function $f(t)$ (roughly) corresponds to multiplication of its transform $\mathcal{L}(f)$ by s, we expect integration of $f(t)$ to correspond to division of $\mathcal{L}(f)$ by s:

THEOREM 3

Laplace Transform of Integral

Let $F(s)$ denote the transform of a function $f(t)$ which is piecewise continuous for $t \geqq 0$ and satisfies a growth restriction (2), Sec. 6.1. Then, for $s > 0$, $s > k$, and $t > 0$,

$$\mathcal{L}\left\{\int_0^t f(\tau)\, d\tau\right\} = \frac{1}{s}F(s), \qquad \text{thus} \qquad \int_0^t f(\tau)\, d\tau = \mathcal{L}^{-1}\left\{\frac{1}{s}F(s)\right\}. \tag{4}$$

PROOF Denote the integral in (4) by $g(t)$. Since $f(t)$ is piecewise continuous, $g(t)$ is continuous, and (2), Sec. 6.1, gives

$$|g(t)| = \left|\int_0^t f(\tau)\, d\tau\right| \leqq \int_0^t |f(\tau)|\, d\tau \leqq M\int_0^t e^{k\tau}\, d\tau = \frac{M}{k}(e^{kt} - 1) \leqq \frac{M}{k}e^{kt} \qquad (k > 0).$$

This shows that $g(t)$ also satisfies a growth restriction. Also, $g'(t) = f(t)$, except at points at which $f(t)$ is discontinuous. Hence $g'(t)$ is piecewise continuous on each finite interval and, by Theorem 1, since $g(0) = 0$ (the integral from 0 to 0 is zero)

$$\mathcal{L}\{f(t)\} = \mathcal{L}\{g'(t)\} = s\mathcal{L}\{g(t)\} - g(0) = s\mathcal{L}\{g(t)\}.$$

Division by s and interchange of the left and right sides gives the first formula in (4), from which the second follows by taking the inverse transform on both sides. ■

EXAMPLE 3 **Application of Theorem 3: Formulas 19 and 20 in the Table of Sec. 6.9**

Using Theorem 3, find the inverse of $\dfrac{1}{s(s^2 + \omega^2)}$ and $\dfrac{1}{s^2(s^2 + \omega^2)}$.

Solution. From Table 6.1 in Sec. 6.1 and the integration in (4) (second formula with the sides interchanged) we obtain

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2 + \omega^2}\right\} = \frac{\sin \omega t}{\omega}, \qquad \mathcal{L}^{-1}\left\{\frac{1}{s(s^2 + \omega^2)}\right\} = \int_0^t \frac{\sin \omega\tau}{\omega}\, d\tau = \frac{1}{\omega^2}(1 - \cos \omega t).$$

This is formula 19 in Sec. 6.9. Integrating this result again and using (4) as before, we obtain formula 20 in Sec. 6.9:

$$\mathcal{L}^{-1}\left\{\frac{1}{s^2(s^2 + \omega^2)}\right\} = \frac{1}{\omega^2}\int_0^t (1 - \cos \omega\tau)\, d\tau = \left[\frac{\tau}{\omega^2} - \frac{\sin \omega\tau}{\omega^3}\right]_0^t = \frac{t}{\omega^2} - \frac{\sin \omega t}{\omega^3}.$$

It is typical that results such as these can be found in several ways. In this example, try partial fraction reduction. ■

Differential Equations, Initial Value Problems

Let us now discuss how the Laplace transform method solves ODEs and initial value problems. We consider an initial value problem

$$y'' + ay' + by = r(t), \qquad y(0) = K_0, \qquad y'(0) = K_1 \tag{5}$$

where a and b are constant. Here $r(t)$ is the given **input** (*driving force*) applied to the mechanical or electrical system and $y(t)$ is the **output** (*response to the input*) to be obtained. In Laplace's method we do three steps:

Step 1. Setting up the subsidiary equation. This is an algebraic equation for the transform $Y = \mathcal{L}(y)$ obtained by transforming (5) by means of (1) and (2), namely,

$$[s^2Y - sy(0) - y'(0)] + a[sY - y(0)] + bY = R(s)$$

where $R(s) = \mathcal{L}(r)$. Collecting the Y-terms, we have the subsidiary equation

$$(s^2 + as + b)Y = (s + a)y(0) + y'(0) + R(s).$$

Step 2. Solution of the subsidiary equation by algebra. We divide by $s^2 + as + b$ and use the so-called **transfer function**

$$Q(s) = \frac{1}{s^2 + as + b} = \frac{1}{(s + \frac{1}{2}a)^2 + b - \frac{1}{4}a^2}. \tag{6}$$

(Q is often denoted by H, but we need H much more frequently for other purposes.) This gives the solution

$$Y(s) = [(s + a)y(0) + y'(0)]Q(s) + R(s)Q(s). \tag{7}$$

If $y(0) = y'(0) = 0$, this is simply $Y = RQ$; hence

$$Q = \frac{Y}{R} = \frac{\mathcal{L}(\text{output})}{\mathcal{L}(\text{input})}$$

and this explains the name of Q. Note that ***Q depends neither on $r(t)$ nor on the initial conditions*** (but only on a and b).

Step 3. Inversion of Y to obtain $y = \mathcal{L}^{-1}(Y)$. We reduce (7) (usually by ***partial fractions*** as in calculus) to a sum of terms whose inverses can be found from the tables (e.g., in Sec. 6.1 or Sec. 6.9) or by a CAS, so that we obtain the solution $y(t) = \mathcal{L}^{-1}(Y)$ *of* (5).

EXAMPLE 4 Initial Value Problem: The Basic Laplace Steps

Solve

$$y'' - y = t, \qquad y(0) = 1, \qquad y'(0) = 1.$$

Solution. ***Step 1.*** From (2) and Table 6.1 we get the subsidiary equation [with $Y = \mathcal{L}(y)$]

$$s^2Y - sy(0) - y'(0) - Y = 1/s^2, \qquad \text{thus} \qquad (s^2 - 1)Y = s + 1 + 1/s^2.$$

Step 2. The transfer function is $Q = 1/(s^2 - 1)$, and (7) becomes

$$Y = (s + 1)Q + \frac{1}{s^2}Q = \frac{s + 1}{s^2 - 1} + \frac{1}{s^2(s^2 - 1)}.$$

Simplification of the first fraction and an expansion of the last fraction gives

$$Y = \frac{1}{s - 1} + \left(\frac{1}{s^2 - 1} - \frac{1}{s^2}\right).$$

Step 3. From this expression for Y and Table 6.1 we obtain the solution

$$y(t) = \mathcal{L}^{-1}(Y) = \mathcal{L}^{-1}\left\{\frac{1}{s-1}\right\} + \mathcal{L}^{-1}\left\{\frac{1}{s^2-1}\right\} - \mathcal{L}^{-1}\left\{\frac{1}{s^2}\right\} = e^t + \sinh t - t.$$

The diagram in Fig. 116 summarizes our approach. ■

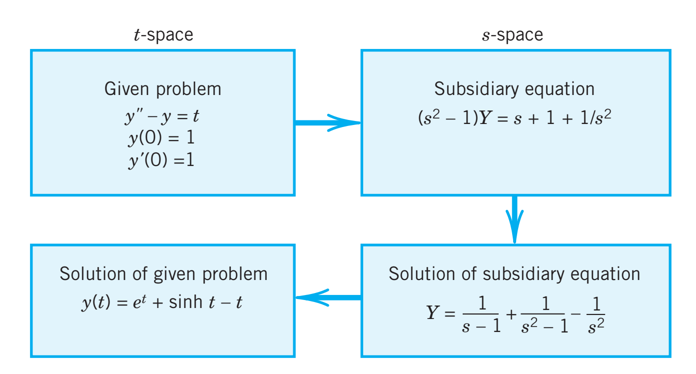


Fig. 116. Steps of the Laplace transform method

EXAMPLE 5 Comparison with the Usual Method

Solve the initial value problem

$$y'' + y' + 9y = 0. \qquad y(0) = 0.16, \qquad y'(0) = 0.$$

Solution. From (1) and (2) we see that the subsidiary equation is

$$s^2Y - 0.16s + sY - 0.16 + 9Y = 0, \qquad \text{thus} \qquad (s^2 + s + 9)Y = 0.16(s + 1).$$

The solution is

$$Y = \frac{0.16(s+1)}{s^2 + s + 9} = \frac{0.16(s + \frac{1}{2}) + 0.08}{(s + \frac{1}{2})^2 + \frac{35}{4}}.$$

Hence by the first shifting theorem and the formulas for cos and sin in Table 6.1 we obtain

$$y(t) = \mathcal{L}^{-1}(Y) = e^{-t/2}\left(0.16 \cos \sqrt{\frac{35}{4}}\, t + \frac{0.08}{\frac{1}{2}\sqrt{35}} \sin \sqrt{\frac{35}{4}}\, t\right)$$

$$= e^{-0.5t}(0.16 \cos 2.96t + 0.027 \sin 2.96t).$$

This agrees with Example 2, Case (III) in Sec. 2.4. The work was less. ■

Advantages of the Laplace Method

1. *Solving a nonhomogeneous ODE does not require first solving the homogeneous ODE.* See Example 4.
2. *Initial values are automatically taken care of.* See Examples 4 and 5.
3. *Complicated inputs* $r(t)$ (right sides of linear ODEs) *can be handled very efficiently,* as we show in the next sections.

EXAMPLE 6 **Shifted Data Problems**

This means initial value problems with initial conditions given at some $t = t_0 > 0$ instead of $t = 0$. For such a problem set $t = \tilde{t} + t_0$, so that $t = t_0$ gives $\tilde{t} = 0$ and the Laplace transform can be applied. For instance, solve

$$y'' + y = 2t, \qquad y(\tfrac{1}{4}\pi) = \tfrac{1}{2}\pi, \qquad y'(\tfrac{1}{4}\pi) = 2 - \sqrt{2}.$$

Solution. We have $t_0 = \frac{1}{4}\pi$ and we set $t = \tilde{t} + \frac{1}{4}\pi$. Then the problem is

$$\tilde{y}'' + \tilde{y} = 2(\tilde{t} + \tfrac{1}{4}\pi), \qquad \tilde{y}(0) = \tfrac{1}{2}\pi, \qquad \tilde{y}'(0) = 2 - \sqrt{2}$$

where $\tilde{y}(\tilde{t}) = y(t)$. Using (2) and Table 6.1 and denoting the transform of $\tilde{y}$ by $\tilde{Y}$, we see that the subsidiary equation of the "shifted" initial value problem is

$$s^2\tilde{Y} - s \cdot \tfrac{1}{2}\pi - (2 - \sqrt{2}) + \tilde{Y} = \frac{2}{s^2} + \frac{\frac{1}{2}\pi}{s}, \qquad \text{thus} \qquad (s^2 + 1)\tilde{Y} = \frac{2}{s^2} + \frac{\frac{1}{2}\pi}{s} + \frac{1}{2}\pi s + 2 - \sqrt{2}.$$

Solving this algebraically for $\tilde{Y}$, we obtain

$$\tilde{Y} = \frac{2}{(s^2 + 1)s^2} + \frac{\frac{1}{2}\pi}{(s^2 + 1)s} + \frac{\frac{1}{2}\pi s}{s^2 + 1} + \frac{2 - \sqrt{2}}{s^2 + 1}.$$

The inverse of the first two terms can be seen from Example 3 (with $\omega = 1$), and the last two terms give cos and sin,

$$\begin{aligned}\tilde{y} = \mathcal{L}^{-1}(\tilde{Y}) &= 2(\tilde{t} - \sin \tilde{t}) + \tfrac{1}{2}\pi(1 - \cos \tilde{t}) + \tfrac{1}{2}\pi \cos \tilde{t} + (2 - \sqrt{2}) \sin \tilde{t} \\ &= 2\tilde{t} + \tfrac{1}{2}\pi - \sqrt{2} \sin \tilde{t}.\end{aligned}$$

Now $\tilde{t} = t - \frac{1}{4}\pi$, $\sin \tilde{t} = \dfrac{1}{\sqrt{2}}(\sin t - \cos t)$, so that the answer (the solution) is

$$y = 2t - \sin t + \cos t.$$

■

PROBLEM SET 6.2

1–11 INITIAL VALUE PROBLEMS (IVPS)

Solve the IVPs by the Laplace transform. If necessary, use partial fraction expansion as in Example 4 of the text. Show all details.

1. $y' + 5.2y = 19.4 \sin 2t, \quad y(0) = 0$

2. $y' + 2y = 0, \quad y(0) = 1.5$

3. $y'' - y' - 6y = 0, \quad y(0) = 11, \quad y'(0) = 28$

4. $y'' + 9y = 10e^{-t}, \quad y(0) = 0, \quad y'(0) = 0$

5. $y'' - \frac{1}{4}y = 0, \quad y(0) = 12, \quad y'(0) = 0$

6. $y'' - 6y' + 5y = 29 \cos 2t, \quad y(0) = 3.2,$ $y'(0) = 6.2$

7. $y'' + 7y' + 12y = 21e^{3t}, \quad y(0) = 3.5,$ $y'(0) = -10$

8. $y'' - 4y' + 4y = 0, \quad y(0) = 8.1, \quad y'(0) = 3.9$

9. $y'' - 4y' + 3y = 6t - 8, \quad y(0) = 0, \quad y'(0) = 0$

10. $y'' + 0.04y = 0.02t^2, \quad y(0) = -25, \quad y'(0) = 0$

11. $y'' + 3y' + 2.25y = 9t^3 + 64, \quad y(0) = 1,$ $y'(0) = 31.5$

12–15 SHIFTED DATA PROBLEMS

Solve the shifted data IVPs by the Laplace transform. Show the details.

12. $y'' - 2y' - 3y = 0, \quad y(4) = -3,$ $y'(4) = -17$

13. $y' - 6y = 0, \quad y(-1) = 4$

14. $y'' + 2y' + 5y = 50t - 100, \quad y(2) = -4,$ $y'(2) = 14$

15. $y'' + 3y' - 4y = 6e^{2t-3}, \quad y(1.5) = 4,$ $y'(1.5) = 5$

16–21 OBTAINING TRANSFORMS BY DIFFERENTIATION

Using (1) or (2), find $\mathcal{L}(f)$ if $f(t)$ equals:

16. $t \cos 4t$

17. te^{-at}

18. $\cos^2 2t$

19. $\sin^2 \omega t$

20. $\sin^4 t$. Use Prob. 19.

21. $\cosh^2 t$

22. PROJECT. Further Results by Differentiation. Proceeding as in Example 1, obtain

(a) $$\mathcal{L}(t \cos \omega t) = \frac{s^2 - \omega^2}{(s^2 + \omega^2)^2}$$

and from this and Example 1: **(b)** formula 21, **(c)** 22, **(d)** 23 in Sec. 6.9,

(e) $$\mathcal{L}(t \cosh at) = \frac{s^2 + a^2}{(s^2 - a^2)^2},$$

(f) $$\mathcal{L}(t \sinh at) = \frac{2as}{(s^2 - a^2)^2}.$$

23–29 INVERSE TRANSFORMS BY INTEGRATION

Using Theorem 3, find $f(t)$ if $\mathcal{L}(F)$ equals:

23. $\dfrac{3}{s^2 + s/4}$

24. $\dfrac{20}{s^3 - 2\pi s^2}$

25. $\dfrac{1}{s(s^2 + \omega^2)}$

26. $\dfrac{1}{s^4 - s^2}$

27. $\dfrac{s + 1}{s^4 + 9s^2}$

28. $\dfrac{3s + 4}{s^4 + k^2 s^2}$

29. $\dfrac{1}{s^3 + as^2}$

30. PROJECT. Comments on Sec. 6.2. (a) Give reasons why Theorems 1 and 2 are more important than Theorem 3.

(b) Extend Theorem 1 by showing that if $f(t)$ is continuous, except for an ordinary discontinuity (finite jump) at some $t = a$ (>0), the other conditions remaining as in Theorem 1, then (see Fig. 117)

(1*) $$\mathcal{L}(f') = s\mathcal{L}(f) - f(0) - [f(a + 0) - f(a - 0)]e^{-as}.$$

(c) Verify (1*) for $f(t) = e^{-t}$ if $0 < t < 1$ and 0 if $t > 1$.

(d) Compare the Laplace transform of solving ODEs with the method in Chap. 2. Give examples of your own to illustrate the advantages of the present method (to the extent we have seen them so far).

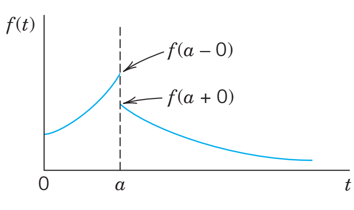


Fig. 117. Formula (1*)

6.3 Unit Step Function (Heaviside Function). Second Shifting Theorem (t-Shifting)

This section and the next one are extremely important because we shall now reach the point where the Laplace transform method shows its real power in applications and its superiority over the classical approach of Chap. 2. The reason is that we shall introduce two auxiliary functions, the *unit step function* or *Heaviside function* $u(t - a)$ (below) and *Dirac's delta* $\delta(t - a)$ (in Sec. 6.4). These functions are suitable for solving ODEs with complicated right sides of considerable engineering interest, such as single waves, inputs (driving forces) that are discontinuous or act for some time only, periodic inputs more general than just cosine and sine, or impulsive forces acting for an instant (hammerblows, for example).

Unit Step Function (Heaviside Function) $u(t - a)$

The **unit step function** or **Heaviside function** $u(t - a)$ is 0 for $t < a$, has a jump of size 1 at $t = a$ (where we can leave it undefined), and is 1 for $t > a$, in a formula:

(1) $$u(t - a) = \begin{cases} 0 & \text{if } t < a \\ 1 & \text{if } t > a \end{cases} \qquad (a \geqq 0).$$

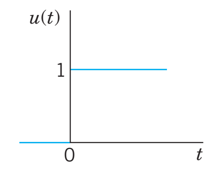


Fig. 118. Unit step function $u(t)$

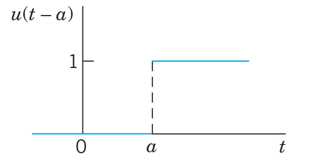


Fig. 119. Unit step function $u(t - a)$

Figure 118 shows the special case $u(t)$, which has its jump at zero, and Fig. 119 the general case $u(t - a)$ for an arbitrary positive a. (For Heaviside, see Sec. 6.1.)

The transform of $u(t - a)$ follows directly from the defining integral in Sec. 6.1,

$$\mathcal{L}\{u(t-a)\} = \int_0^\infty e^{-st}u(t-a)\,dt = \int_0^\infty e^{-st} \cdot 1\,dt = -\frac{e^{-st}}{s}\bigg|_{t=a}^{\infty};$$

here the integration begins at $t = a\,(\geqq 0)$ because $u(t - a)$ is 0 for $t < a$. Hence

$$\mathcal{L}\{u(t-a)\} = \frac{e^{-as}}{s} \qquad (s > 0). \tag{2}$$

The unit step function is a typical "engineering function" made to measure for engineering applications, which often involve functions (mechanical or electrical driving forces) that are either "off" or "on." Multiplying functions $f(t)$ with $u(t - a)$, we can produce all sorts of effects. The simple basic idea is illustrated in Figs. 120 and 121. In Fig. 120 the given function is shown in (A). In (B) it is switched off between $t = 0$ and $t = 2$ (because $u(t - 2) = 0$ when $t < 2$) and is switched on beginning at $t = 2$. In (C) it is shifted to the right by 2 units, say, for instance, by 2 sec, so that it begins 2 sec later in the same fashion as before. More generally we have the following.

Let $f(t) = 0$ for all negative t. Then $f(t - a)u(t - a)$ with $a > 0$ is $f(t)$ ***shifted*** *(translated) to the right by the amount a.*

Figure 121 shows the effect of many unit step functions, three of them in (A) and infinitely many in (B) when continued periodically to the right; this is the effect of a rectifier that clips off the negative half-waves of a sinuosidal voltage. **CAUTION!** Make sure that you fully understand these figures, in particular the difference between parts (B) and (C) of Fig. 120. Figure 120(C) will be applied next.

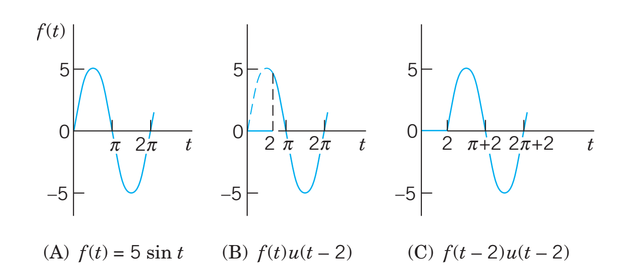
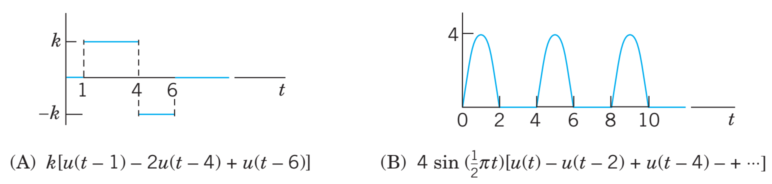


Fig. 120. Effects of the unit step function: (A) Given function. (B) Switching off and on. (C) Shift.



(A) $k[u(t-1) - 2u(t-4) + u(t-6)]$ (B) $4 \sin (\frac{1}{2}\pi t)[u(t) - u(t-2) + u(t-4) - + \cdots]$

Fig. 121. Use of many unit step functions.

Time Shifting (t-Shifting): Replacing t by $t - a$ in $f(t)$

The first shifting theorem ("s-shifting") in Sec. 6.1 concerned transforms $F(s) = \mathcal{L}\{f(t)\}$ and $F(s - a) = \mathcal{L}\{e^{at}f(t)\}$. The second shifting theorem will concern functions $f(t)$ and $f(t - a)$. Unit step functions are just tools, and the theorem will be needed to apply them in connection with any other functions.

THEOREM 1

Second Shifting Theorem; Time Shifting

If $f(t)$ has the transform $F(s)$, then the "**shifted function**"

$$\tilde{f}(t) = f(t-a)u(t-a) = \begin{cases} 0 & \text{if } t < a \\ f(t-a) & \text{if } t > a \end{cases} \tag{3}$$

has the transform $e^{-as}F(s)$. *That is, if* $\mathcal{L}\{f(t)\} = F(s)$, *then*

$$\mathcal{L}\{f(t-a)u(t-a)\} = e^{-as}F(s). \tag{4}$$

Or, if we take the inverse on both sides, we can write

$$f(t-a)u(t-a) = \mathcal{L}^{-1}\{e^{-as}F(s)\}. \tag{4*}$$

Practically speaking, if we know $F(s)$, we can obtain the transform of (3) by multiplying $F(s)$ by e^{-as}. In Fig. 120, the transform of $5 \sin t$ is $F(s) = 5/(s^2 + 1)$, hence the shifted function $5 \sin (t-2)u(t-2)$ shown in Fig. 120(C) has the transform

$$e^{-2s}F(s) = 5e^{-2s}/(s^2 + 1).$$

PROOF We prove Theorem 1. In (4), on the right, we use the definition of the Laplace transform, writing τ for t (to have t available later). Then, taking e^{-as} inside the integral, we have

$$e^{-as}F(s) = e^{-as}\int_0^\infty e^{-s\tau}f(\tau)\,d\tau = \int_0^\infty e^{-s(\tau+a)}f(\tau)\,d\tau.$$

Substituting $\tau + a = t$, thus $\tau = t - a$, $d\tau = dt$ in the integral (**CAUTION**, the lower limit changes!), we obtain

$$e^{-as}F(s) = \int_a^\infty e^{-st}f(t-a)\,dt.$$

To make the right side into a Laplace transform, we must have an integral from 0 to ∞, not from a to ∞. But this is easy. We multiply the integrand by $u(t-a)$. Then for t from 0 to a the integrand is 0, and we can write, with $\tilde{f}$ as in (3),

$$e^{-as}F(s) = \int_0^\infty e^{-st} f(t-a)u(t-a)\,dt = \int_0^\infty e^{-st}\tilde{f}(t)\,dt.$$

(Do you now see why $u(t-a)$ appears?) This integral is the left side of (4), the Laplace transform of $\tilde{f}(t)$ in (3). This completes the proof. ■

EXAMPLE 1 Application of Theorem 1. Use of Unit Step Functions

Write the following function using unit step functions and find its transform.

$$f(t) = \begin{cases} 2 & \text{if } 0 < t < 1 \\ \frac{1}{2}t^2 & \text{if } 1 < t < \frac{1}{2}\pi \\ \cos t & \text{if } \quad t > \frac{1}{2}\pi. \end{cases} \qquad \text{(Fig. 122)}$$

Solution. ***Step 1.*** In terms of unit step functions,

$$f(t) = 2(1 - u(t-1)) + \tfrac{1}{2}t^2(u(t-1) - u(t-\tfrac{1}{2}\pi)) + (\cos t)u(t-\tfrac{1}{2}\pi).$$

Indeed, $2(1-u(t-1))$ gives $f(t)$ for $0 < t < 1$, and so on.

Step 2. To apply Theorem 1, we must write each term in $f(t)$ in the form $f(t-a)u(t-a)$. Thus, $2(1-u(t-1))$ remains as it is and gives the transform $2(1-e^{-s})/s$. Then

$$\mathcal{L}\left\{\frac{1}{2}t^2u(t-1)\right\} = \mathcal{L}\left(\frac{1}{2}(t-1)^2 + (t-1) + \frac{1}{2}\right)u(t-1)\right\} = \left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)e^{-s}$$

$$\mathcal{L}\left\{\frac{1}{2}t^2u\left(t-\frac{1}{2}\pi\right)\right\} = \mathcal{L}\left\{\frac{1}{2}\left(t-\frac{1}{2}\pi\right)^2 + \frac{\pi}{2}\left(t-\frac{1}{2}\pi\right) + \frac{\pi^2}{8}\right)u\left(t-\frac{1}{2}\pi\right)\right\}$$

$$= \left(\frac{1}{s^3} + \frac{\pi}{2s^2} + \frac{\pi^2}{8s}\right)e^{-\pi s/2}$$

$$\mathcal{L}\left\{(\cos t)\,u\left(t-\frac{1}{2}\pi\right)\right\} = \mathcal{L}\left\{-\left(\sin\left(t-\frac{1}{2}\pi\right)\right)u\left(t-\frac{1}{2}\pi\right)\right\} = -\frac{1}{s^2+1}e^{-\pi s/2}.$$

Together,

$$\mathcal{L}(f) = \frac{2}{s} - \frac{2}{s}e^{-s} + \left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)e^{-s} - \left(\frac{1}{s^3} + \frac{\pi}{2s^2} + \frac{\pi^2}{8s}\right)e^{-\pi s/2} - \frac{1}{s^2+1}e^{-\pi s/2}.$$

If the conversion of $f(t)$ to $f(t-a)$ is inconvenient, replace it by

(4)** $$\mathcal{L}\{f(t)u(t-a)\} = e^{-as}\mathcal{L}\{f(t+a)\}.$$

(4**) follows from (4) by writing $f(t-a) = g(t)$, hence $f(t) = g(t+a)$ and then again writing f for g. Thus,

$$\mathcal{L}\left\{\frac{1}{2}t^2u(t-1)\right\} = e^{-s}\mathcal{L}\left\{\frac{1}{2}(t+1)^2\right\} = e^{-s}\mathcal{L}\left\{\frac{1}{2}t^2 + t + \frac{1}{2}\right\} = e^{-s}\left(\frac{1}{s^3} + \frac{1}{s^2} + \frac{1}{2s}\right)$$

as before. Similarly for $\mathcal{L}\{\frac{1}{2}t^2u(t-\frac{1}{2}\pi)\}$. Finally, by (4**),

$$\mathcal{L}\left\{\cos t\,u\left(t-\frac{1}{2}\pi\right)\right\} = e^{-\pi s/2}\mathcal{L}\left\{\cos\left(t+\frac{1}{2}\pi\right)\right\} = e^{-\pi s/2}\mathcal{L}\{-\sin t\} = -e^{-\pi s/2}\frac{1}{s^2+1}.$$ ■

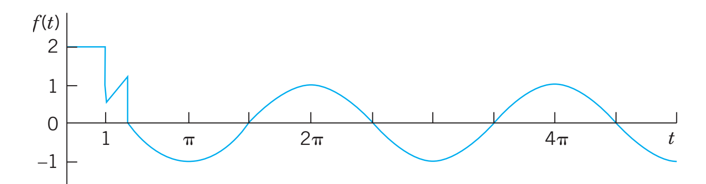


Fig. 122. $f(t)$ in Example 1

EXAMPLE 2 Application of Both Shifting Theorems. Inverse Transform

Find the inverse transform $f(t)$ of

$$F(s) = \frac{e^{-s}}{s^2 + \pi^2} + \frac{e^{-2s}}{s^2 + \pi^2} + \frac{e^{-3s}}{(s+2)^2}.$$

Solution. Without the exponential functions in the numerator the three terms of $F(s)$ would have the inverses $(\sin \pi t)/\pi$, $(\sin \pi t)/\pi$, and te^{-2t} because $1/s^2$ has the inverse t, so that $1/(s+2)^2$ has the inverse te^{-2t} by the first shifting theorem in Sec. 6.1. Hence by the second shifting theorem (t-shifting),

$$f(t) = \frac{1}{\pi} \sin(\pi(t-1))\, u(t-1) + \frac{1}{\pi} \sin(\pi(t-2))\, u(t-2) + (t-3)e^{-2(t-3)}\, u(t-3).$$

Now $\sin(\pi t - \pi) = -\sin \pi t$ and $\sin(\pi t - 2\pi) = \sin \pi t$, so that the first and second terms cancel each other when $t > 2$. Hence we obtain $f(t) = 0$ if $0 < t < 1$, $-(\sin \pi t)/\pi$ if $1 < t < 2$, 0 if $2 < t < 3$, and $(t-3)e^{-2(t-3)}$ if $t > 3$. See Fig. 123. ■

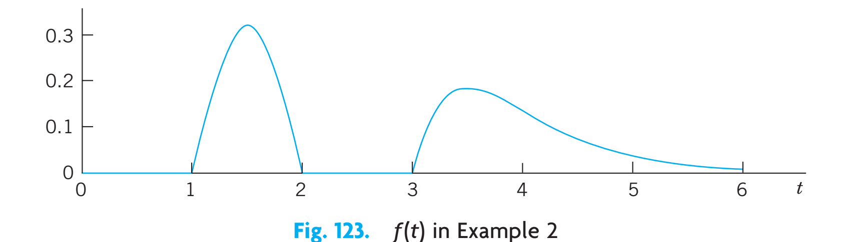


Fig. 123. $f(t)$ in Example 2

EXAMPLE 3 Response of an *RC*-Circuit to a Single Rectangular Wave

Find the current $i(t)$ in the RC-circuit in Fig. 124 if a single rectangular wave with voltage V_0 is applied. The circuit is assumed to be quiescent before the wave is applied.

Solution. The input is $V_0[u(t-a) - u(t-b)]$. Hence the circuit is modeled by the integro-differential equation (see Sec. 2.9 and Fig. 124)

$$Ri(t) + \frac{q(t)}{C} = Ri(t) + \frac{1}{C}\int_0^t i(\tau)\, d\tau = v(t) = V_0[u(t-a) - u(t-b)].$$

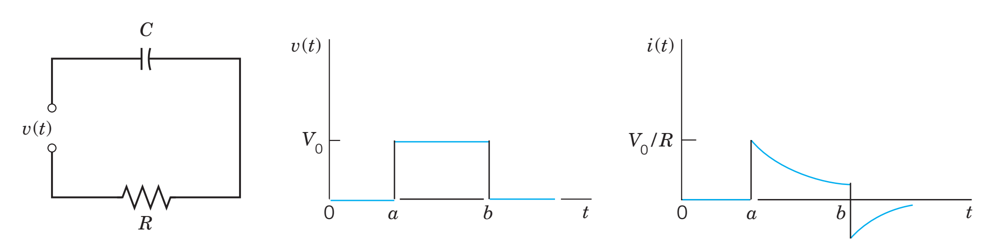


Fig. 124. RC-circuit, electromotive force $v(t)$, and current in Example 3

Using Theorem 3 in Sec. 6.2 and formula (1) in this section, we obtain the subsidiary equation

$$RI(s) + \frac{I(s)}{sC} = \frac{V_0}{s}[e^{-as} - e^{-bs}].$$

Solving this equation algebraically for $I(s)$, we get

$$I(s) = F(s)(e^{-as} - e^{-bs}) \qquad \text{where} \qquad F(s) = \frac{V_0 IR}{s + 1/(RC)} \qquad \text{and} \qquad \mathcal{L}^{-1}(F) = \frac{V_0}{R}e^{-t/(RC)},$$

the last expression being obtained from Table 6.1 in Sec. 6.1. Hence Theorem 1 yields the solution (Fig. 124)

$$i(t) = \mathcal{L}^{-1}(I) = \mathcal{L}^{-1}\{e^{-as}F(s)\} - \mathcal{L}^{-1}\{e^{-bs}F(s)\} = \frac{V_0}{R}[e^{-(t-a)/(RC)}u(t-a) - e^{-(t-b)/(RC)}u(t-b)];$$

that is, $i(t) = 0$ if $t < a$, and

$$i(t) = \begin{cases} K_1 e^{-t/(RC)} & \text{if } a < t < b \\ (K_1 - K_2)e^{-t/(RC)} & \text{if } a > b \end{cases}$$

where $K_1 = V_0 e^{a/(RC)}/R$ and $K_2 = V_0 e^{b/(RC)}/R$. ■

EXAMPLE 4 Response of an *RLC*-Circuit to a Sinusoidal Input Acting Over a Time Interval

Find the response (the current) of the *RLC*-circuit in Fig. 125, where $E(t)$ is sinusoidal, acting for a short time interval only, say,

$$E(t) = 100 \sin 400t \quad \text{if } 0 < t < 2\pi \qquad \text{and} \qquad E(t) = 0 \text{ if } t > 2\pi$$

and current and charge are initially zero.

Solution. The electromotive force $E(t)$ can be represented by $(100 \sin 400t)(1 - u(t - 2\pi))$. Hence the model for the current $i(t)$ in the circuit is the integro-differential equation (see Sec. 2.9)

$$0.1i' + 11i + 100\int_0^t i(\tau)\,d\tau = (100 \sin 400t)(1 - u(t - 2\pi)). \qquad i(0) = 0, \quad i'(0) = 0.$$

From Theorems 2 and 3 in Sec. 6.2 we obtain the subsidiary equation for $I(s) = \mathcal{L}(i)$

$$0.1sI + 11I + 100\frac{I}{s} = \frac{100 \cdot 400s}{s^2 + 400^2}\left(\frac{1}{s} - \frac{e^{-2\pi s}}{s}\right).$$

Solving it algebraically and noting that $s^2 + 110s + 1000 = (s + 10)(s + 100)$, we obtain

$$I(s) = \frac{1000 \cdot 400}{(s+10)(s+100)}\left(\frac{s}{s^2 + 400^2} - \frac{se^{-2\pi s}}{s^2 + 400^2}\right).$$

For the first term in the parentheses $(\cdots)$ times the factor in front of them we use the partial fraction expansion

$$\frac{400{,}000s}{(s+10)(s+100)(s^2 + 400^2)} = \frac{A}{s+10} + \frac{B}{s+100} + \frac{Ds + K}{s^2 + 400^2}.$$

Now determine A, B, D, K by your favorite method or by a CAS or as follows. Multiplication by the common denominator gives

$$400{,}000s = A(s + 100)(s^2 + 400^2) + B(s + 10)(s^2 + 400^2) + (Ds + K)(s + 10)(s + 100).$$

We set $s = -10$ and -100 and then equate the sums of the s^3 and s^2 terms to zero, obtaining (all values rounded)

$$(s = -10) \qquad -4{,}000{,}000 = 90(10^2 + 400^2)A, \qquad A = -0.27760$$

$$(s = -100) \qquad -40{,}000{,}000 = -90(100^2 + 400^2)B, \qquad B = 2.6144$$

$$(s^3\text{-terms}) \qquad 0 = A + B + D, \qquad D = -2.3368$$

$$(s^2\text{-terms}) \qquad 0 = 100A + 10B + 110D + K, \qquad K = 258.66.$$

Since $K = 258.66 = 0.6467 \cdot 400$, we thus obtain for the first term I_1 in $I = I_1 - I_2$

$$I_1 = -\frac{0.2776}{s + 10} + \frac{2.6144}{s + 100} - \frac{2.3368s}{s^2 + 400^2} + \frac{0.6467 \cdot 400}{s^2 + 400^2}.$$

From Table 6.1 in Sec. 6.1 we see that its inverse is

$$i_1(t) = -0.2776e^{-10t} + 2.6144e^{-100t} - 2.3368 \cos 400t + 0.6467 \sin 400t.$$

This is the current $i(t)$ when $0 < t < 2\pi$. It agrees for $0 < t < 2\pi$ with that in Example 1 of Sec. 2.9 (except for notation), which concerned the same RLC-circuit. Its graph in Fig. 63 in Sec. 2.9 shows that the exponential terms decrease very rapidly. Note that the present amount of work was substantially less.

The second term I_1 of I differs from the first term by the factor $e^{-2\pi s}$. Since $\cos 400(t - 2\pi) = \cos 400t$ and $\sin 400(t - 2\pi) = \sin 400t$, the second shifting theorem (Theorem 1) gives the inverse $i_2(t) = 0$ if $0 < t < 2\pi$, and for $> 2\pi$ it gives

$$i_2(t) = -0.2776e^{-10(t-2\pi)} + 2.6144e^{-100(t-2\pi)} - 2.3368 \cos 400t + 0.6467 \sin 400t.$$

Hence in $i(t)$ the cosine and sine terms cancel, and the current for $t > 2\pi$ is

$$i(t) = -0.2776(e^{-10t} - e^{-10(t-2\pi)}) + 2.6144(e^{-100t} - e^{-100(t-2\pi)}).$$

It goes to zero very rapidly, practically within 0.5 sec. ■

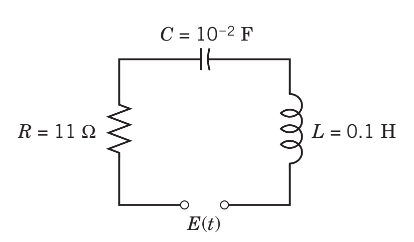


Fig. 125. *RLC*-circuit in Example 4

PROBLEM SET 6.3

1. Report on Shifting Theorems. Explain and compare the different roles of the two shifting theorems, using your own formulations and simple examples. Give no proofs.

2–11 SECOND SHIFTING THEOREM, UNIT STEP FUNCTION

Sketch or graph the given function, which is assumed to be zero outside the given interval. Represent it, using unit step functions. Find its transform. Show the details of your work.

2. $t \ (0 < t < 2)$ **3.** $t - 2 \ (t > 2)$

4. $\cos 4t \ (0 < t < \pi)$ **5.** $e^t \ (0 < t < \pi/2)$

6. $\sin \pi t \ (2 < t < 4)$ **7.** $e^{-\pi t} \ (2 < t < 4)$

8. $t^2 \ (1 < t < 2)$ **9.** $t^2 \ (t > \frac{3}{2})$

10. $\sinh t \ (0 < t < 2)$ **11.** $\sin t \ (\pi/2 < t < \pi)$

12–17 INVERSE TRANSFORMS BY THE 2ND SHIFTING THEOREM

Find and sketch or graph $f(t)$ if $\mathcal{L}(f)$ equals

12. $e^{-3s}/(s - 1)^3$ **13.** $6(1 - e^{-\pi s})/(s^2 + 9)$

14. $4(e^{-2s} - 2e^{-5s})/s$ **15.** e^{-3s}/s^4

16. $2(e^{-s} - e^{-3s})/(s^2 - 4)$

17. $(1 + e^{-2\pi(s+1)})(s + 1)/((s + 1)^2 + 1)$

18–27 IVPs, SOME WITH DISCONTINUOUS INPUT

Using the Laplace transform and showing the details, solve

18. $9y'' - 6y' + y = 0, \quad y(0) = 3, y'(0) = 1$

19. $y'' + 6y' + 8y = e^{-3t} - e^{-5t}, \quad y(0) = 0, y'(0) = 0$

20. $y'' + 10y' + 24y = 144t^2, \quad y(0) = 19/12, y'(0) = -5$

21. $y'' + 9y = 8 \sin t$ if $0 < t < \pi$ and 0 if $t > \pi$; $y(0) = 0, y'(0) = 4$

22. $y'' + 3y' + 2y = 4t$ if $0 < t < 1$ and 8 if $t > 1$; $y(0) = 0, y'(0) = 0$

23. $y'' + y' - 2y = 3 \sin t - \cos t$ if $0 < t < 2\pi$ and $3 \sin 2t - \cos 2t$ if $t > 2\pi$; $y(0) = 1, y'(0) = 0$

24. $y'' + 3y' + 2y = 1$ if $0 < t < 1$ and 0 if $t > 1$; $y(0) = 0, y'(0) = 0$

25. $y'' + y = t$ if $0 < t < 1$ and 0 if $t > 1$; $y(0) = 0, y'(0) = 0$

26. Shifted data. $y'' + 2y' + 5y = 10 \sin t$ if $0 < t < 2\pi$ and 0 if $t > 2\pi$; $y(\pi) = 1, y'(\pi) = 2e^{-\pi} - 2$

27. Shifted data. $y'' + 4y = 8t^2$ if $0 < t < 5$ and 0 if $t > 5$; $y(1) = 1 + \cos 2, y'(1) = 4 - 2 \sin 2$

28–40 MODELS OF ELECTRIC CIRCUITS

28–30 *RL*-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 126, assuming $i(0) = 0$ and:

28. $R = 1$ kΩ (=1000 Ω), $L = 1$ H, $v = 0$ if $0 < t < \pi$, and $40 \sin t$ V if $t > \pi$

29. $R = 25$ Ω, $L = 0.1$ H, $v = 490\, e^{-5t}$ V if $0 < t < 1$ and 0 if $t > 1$

30. $R = 10$ Ω, $L = 0.5$ H, $v = 200t$ V if $0 < t < 2$ and 0 if $t > 2$

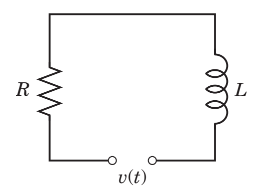


Fig. 126. Problems 28–30

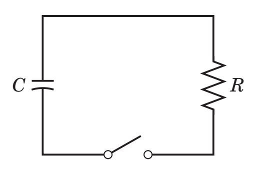


Fig. 127. Problem 31

31. Discharge in *RC*-circuit. Using the Laplace transform, find the charge $q(t)$ on the capacitor of capacitance C in Fig. 127 if the capacitor is charged so that its potential is V_0 and the switch is closed at $t = 0$.

32–34 *RC*-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 128 with $R = 10$ Ω and $C = 10^{-2}$ F, where the current at $t = 0$ is assumed to be zero, and:

32. $v = 0$ if $t < 4$ and $14 \cdot 10^6 e^{-3t}$ V if $t > 4$

33. $v = 0$ if $t < 2$ and $100(t - 2)$ V if $t > 2$

34. $v(t) = 100$ V if $0.5 < t < 0.6$ and 0 otherwise. Why does $i(t)$ have jumps?

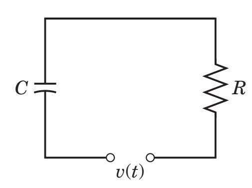


Fig. 128. Problems 32–34

35–37 *LC*-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 129, assuming zero initial current and charge on the capacitor and:

35. $L = 1$ H, $C = 10^{-2}$ F, $v = -9900 \cos t$ V if $\pi < t < 3\pi$ and 0 otherwise

36. $L = 1$ H, $C = 0.25$ F, $v = 200\,(t - \frac{1}{3}t^3)$ V if $0 < t < 1$ and 0 if $t > 1$

37. $L = 0.5$ H, $C = 0.05$ F, $v = 78 \sin t$ V if $0 < t < \pi$ and 0 if $t > \pi$

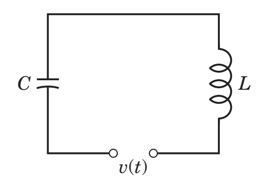


Fig. 129. Problems 35–37

38–40 *RLC*-CIRCUIT

Using the Laplace transform and showing the details, find the current $i(t)$ in the circuit in Fig. 130, assuming zero initial current and charge and:

38. $R = 4$ Ω, $L = 1$ H, $C = 0.05$ F, $v = 34e^{-t}$ V if $0 < t < 4$ and 0 if $t > 4$

39. $R = 2\ \Omega$, $L = 1$ H, $C = 0.5$ F, $v(t) = 1$ kV if $0 < t < 2$ and 0 if $t > 2$

40. $R = 2\ \Omega$, $L = 1$ H, $C = 0.1$ F, $v = 255 \sin t$ V if $0 < t < 2\pi$ and 0 if $t > 2\pi$

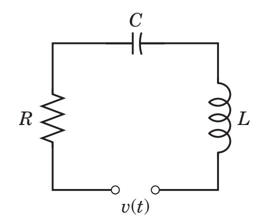


Fig. 130. Problems 38–40

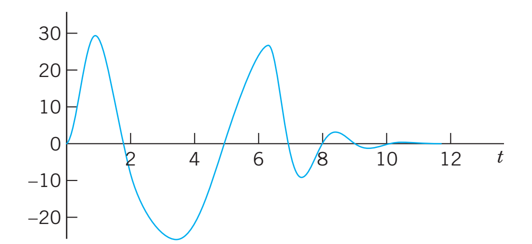


Fig. 131. Current in Problem 40

6.4 Short Impulses. Dirac's Delta Function. Partial Fractions

An airplane making a "hard" landing, a mechanical system being hit by a hammerblow, a ship being hit by a single high wave, a tennis ball being hit by a racket, and many other similar examples appear in everyday life. They are phenomena of an impulsive nature where actions of forces—mechanical, electrical, etc.—are applied over short intervals of time.

We can model such phenomena and problems by "Dirac's delta function," and solve them very effecively by the Laplace transform.

To model situations of that type, we consider the function

$$f_k(t-a) = \begin{cases} 1/k & \text{if } a \leqq t \leqq a+k \\ 0 & \text{otherwise} \end{cases} \tag{1}$$ (Fig. 132)

(and later its limit as $k \to 0$). This function represents, for instance, a force of magnitude $1/k$ acting from $t = a$ to $t = a + k$, where k is positive and small. In mechanics, the integral of a force acting over a time interval $a \leqq t \leqq a + k$ is called the **impulse** of the force; similarly for electromotive forces $E(t)$ acting on circuits. Since the blue rectangle in Fig. 132 has area 1, the impulse of f_k in (1) is

$$I_k = \int_0^\infty f_k(t-a)\, dt = \int_a^{a+k} \frac{1}{k}\, dt = 1. \tag{2}$$

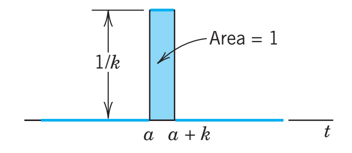


Fig. 132. The function $f_k(t - a)$ in (1)

To find out what will happen if k becomes smaller and smaller, we take the limit of f_k as $k \to 0$ $(k > 0)$. This limit is denoted by $\delta(t - a)$, that is,

$$\delta(t-a) = \lim_{k \to 0} f_k(t-a).$$

$\delta(t - a)$ is called the **Dirac delta function**[2] or the **unit impulse function**.

$\delta(t - a)$ is not a function in the ordinary sense as used in calculus, but a so-called *generalized function*.[2] To see this, we note that the impulse I_k of f_k is 1, so that from (1) and (2) by taking the limit as $k \to 0$ we obtain

$$(3) \qquad \delta(t-a) = \begin{cases} \infty & \text{if } t = a \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \int_0^\infty \delta(t-a)\,dt = 1,$$

but from calculus we know that a function which is everywhere 0 except at a single point must have the integral equal to 0. Nevertheless, in impulse problems, it is convenient to operate on $\delta(t - a)$ as though it were an ordinary function. In particular, for a *continuous* function $g(t)$ one uses the property [often called the **sifting property** of $\delta(t - a)$, not to be confused with *shifting*]

$$\textbf{(4)} \qquad \int_0^\infty g(t)\delta(t-a)\,dt = g(a)$$

which is plausible by (2).

To obtain the Laplace transform of $\delta(t - a)$, we write

$$f_k(t-a) = \frac{1}{k}\left[u(t-a) - u(t-(a+k))\right]$$

and take the transform [see (2)]

$$\mathcal{L}\{f_k(t-a)\} = \frac{1}{ks}\left[e^{-as} - e^{-(a+k)s}\right] = e^{-as}\frac{1-e^{-ks}}{ks}.$$

We now take the limit as $k \to 0$. By l'Hôpital's rule the quotient on the right has the limit 1 (differentiate the numerator and the denominator separately with respect to k, obtaining se^{-ks} and s, respectively, and use $se^{-ks}/s \to 1$ as $k \to 0$). Hence the right side has the limit e^{-as}. This suggests defining the transform of $\delta(t - a)$ by this limit, that is,

$$\textbf{(5)} \qquad \mathcal{L}\{\delta(t-a)\} = e^{-as}.$$

The unit step and unit impulse functions can now be used on the right side of ODEs modeling mechanical or electrical systems, as we illustrate next.

---

[2]PAUL DIRAC (1902–1984), English physicist, was awarded the Nobel Prize [jointly with the Austrian ERWIN SCHRÖDINGER (1887–1961)] in 1933 for his work in quantum mechanics.

Generalized functions are also called **distributions**. Their theory was created in 1936 by the Russian mathematician SERGEI L'VOVICH SOBOLEV (1908–1989), and in 1945, under wider aspects, by the French mathematician LAURENT SCHWARTZ (1915–2002).

EXAMPLE 1 **Mass–Spring System Under a Square Wave**

Determine the response of the damped mass–spring system (see Sec. 2.8) under a square wave, modeled by (see Fig. 133)

$$y'' + 3y' + 2y = r(t) = u(t-1) - u(t-2), \qquad y(0) = 0, \qquad y'(0) = 0.$$

Solution. From (1) and (2) in Sec. 6.2 and (2) and (4) in this section we obtain the subsidiary equation

$$s^2Y + 3sY + 2Y = \frac{1}{s}(e^{-s} - e^{-2s}). \qquad \text{Solution} \qquad Y(s) = \frac{1}{s(s^2+3s+2)}(e^{-s} - e^{-2s}).$$

Using the notation $F(s)$ and partial fractions, we obtain

$$F(s) = \frac{1}{s(s^2+3s+2)} = \frac{1}{s(s+1)(s+2)} = \frac{\frac{1}{2}}{s} - \frac{1}{s+1} + \frac{\frac{1}{2}}{s+2}.$$

From Table 6.1 in Sec. 6.1, we see that the inverse is

$$f(t) = \mathcal{L}^{-1}(F) = \tfrac{1}{2} - e^{-t} + \tfrac{1}{2}e^{-2t}.$$

Therefore, by Theorem 1 in Sec. 6.3 (t-shifting) we obtain the square-wave response shown in Fig. 133,

$$\begin{aligned} y &= \mathcal{L}^{-1}(F(s)e^{-s} - F(s)e^{-2s}) \\ &= f(t-1)u(t-1) - f(t-2)u(t-2) \\ &= \begin{cases} 0 & (0 < t < 1) \\ \frac{1}{2} - e^{-(t-1)} + \frac{1}{2}e^{-2(t-1)} & (1 < t < 2) \\ -e^{-(t-1)} + e^{-(t-2)} + \frac{1}{2}e^{-2(t-1)} - \frac{1}{2}e^{-2(t-2)} & (t > 2). \end{cases} \end{aligned}$$

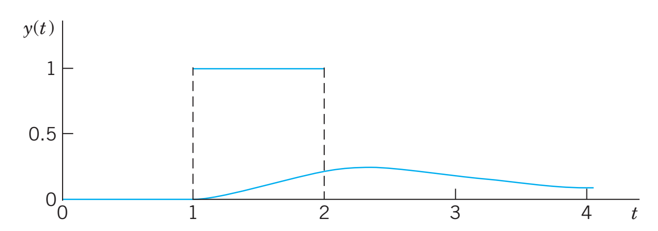


Fig. 133. Square wave and response in Example 1

EXAMPLE 2 **Hammerblow Response of a Mass–Spring System**

Find the response of the system in Example 1 with the square wave replaced by a unit impulse at time $t = 1$.

Solution. We now have the ODE and the subsidiary equation

$$y'' + 3y' + 2y = \delta(t-1), \qquad \text{and} \qquad (s^2 + 3s + 2)Y = e^{-s}.$$

Solving algebraically gives

$$Y(s) = \frac{e^{-s}}{(s+1)(s+2)} = \left(\frac{1}{s+1} - \frac{1}{s+2}\right)e^{-s}.$$

By Theorem 1 the inverse is

$$y(t) = \mathcal{L}^{-1}(Y) = \begin{cases} 0 & \text{if } 0 < t < 1 \\ e^{-(t-1)} - e^{-2(t-1)} & \text{if } \quad t > 1. \end{cases}$$

$y(t)$ is shown in Fig. 134. Can you imagine how Fig. 133 approaches Fig. 134 as the wave becomes shorter and shorter, the area of the rectangle remaining 1? ■

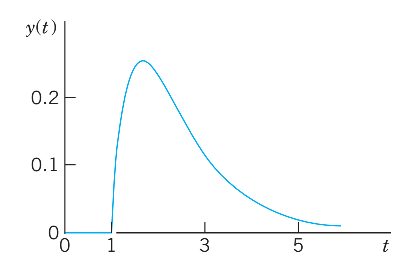


Fig. 134. Response to a hammerblow in Example 2

EXAMPLE 3 Four-Terminal *RLC*-Network

Find the output voltage response in Fig. 135 if $R = 20\ \Omega$, $L = 1$ H, $C = 10^{-4}$ F, the input is $\delta(t)$ (a unit impulse at time $t = 0$), and current and charge are zero at time $t = 0$.

Solution. To understand what is going on, note that the network is an *RLC*-circuit to which two wires at A and B are attached for recording the voltage $v(t)$ on the capacitor. Recalling from Sec. 2.9 that current $i(t)$ and charge $q(t)$ are related by $i = q' = dq/dt$, we obtain the model

$$Li' + Ri + \frac{q}{C} = Lq'' + Rq' + \frac{q}{C} = q'' + 20q' + 10{,}000q = \delta(t).$$

From (1) and (2) in Sec. 6.2 and (5) in this section we obtain the subsidiary equation for $Q(s) = \mathcal{L}(q)$

$$(s^2 + 20s + 10{,}000)Q = 1. \qquad \text{Solution} \qquad Q = \frac{1}{(s + 10)^2 + 9900}.$$

By the first shifting theorem in Sec. 6.1 we obtain from Q damped oscillations for q and v; rounding $9900 \approx 99.50^2$, we get (Fig. 135)

$$q = \mathcal{L}^{-1}(Q) = \frac{1}{99.50} e^{-10t} \sin 99.50t \qquad \text{and} \qquad v = \frac{q}{C} = 100.5e^{-10t} \sin 99.50t. \qquad ■$$

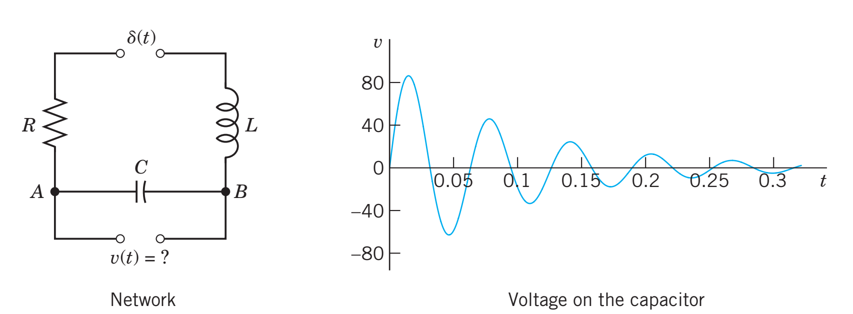


Fig. 135. Network and output voltage in Example 3

More on Partial Fractions

We have seen that the solution Y of a subsidiary equation usually appears as a quotient of polynomials $Y(s) = F(s)/G(s)$, so that a partial fraction representation leads to a sum of expressions whose inverses we can obtain from a table, aided by the first shifting theorem (Sec. 6.1). These representations are sometimes called **Heaviside expansions**.

An *unrepeated factor* $s - a$ in $G(s)$ requires a single partial fraction $A/(s - a)$. See Examples 1 and 2. *Repeated real factors* $(s - a)^2$, $(s - a)^3$, etc., require partial fractions

$$\frac{A_2}{(s-a)^2} + \frac{A_1}{s-a}, \qquad \frac{A_3}{(s-a)^3} + \frac{A_2}{(s-a)^2} + \frac{A_1}{s-a}, \quad \text{etc.,}$$

The inverses are $(A_2 t + A_1)e^{at}$, $(\frac{1}{2}A_3 t^2 + A_2 t + A_1)e^{at}$, etc.

Unrepeated complex factors $(s - a)(s - \bar{a})$, $a = \alpha + i\beta$, $\bar{a} = \alpha - i\beta$, require a partial fraction $(As + B)/[(s - \alpha)^2 + \beta^2]$. For an application, see Example 4 in Sec. 6.3. A further one is the following.

EXAMPLE 4 Unrepeated Complex Factors. Damped Forced Vibrations

Solve the initial value problem for a damped mass–spring system acted upon by a sinusoidal force for some time interval (Fig. 136),

$$y'' + 2y' + 2y = r(t), \quad r(t) = 10 \sin 2t \text{ if } 0 < t < \pi \text{ and } 0 \text{ if } t > \pi; \qquad y(0) = 1, \quad y'(0) = -5.$$

Solution. From Table 6.1, (1), (2) in Sec. 6.2, and the second shifting theorem in Sec. 6.3, we obtain the subsidiary equation

$$(s^2Y - s + 5) + 2(sY - 1) + 2Y = 10\,\frac{2}{s^2+4}(1 - e^{-\pi s}).$$

We collect the Y-terms, $(s^2 + 2s + 2)Y$, take $-s + 5 - 2 = -s + 3$ to the right, and solve,

$$Y = \frac{20}{(s^2+4)(s^2+2s+2)} - \frac{20e^{-\pi s}}{(s^2+4)(s^2+2s+2)} + \frac{s-3}{s^2+2s+2}. \tag{6}$$

For the last fraction we get from Table 6.1 and the first shifting theorem

$$\mathcal{L}^{-1}\left\{\frac{s+1-4}{(s+1)^2+1}\right\} = e^{-t}(\cos t - 4 \sin t). \tag{7}$$

In the first fraction in (6) we have unrepeated complex roots, hence a partial fraction representation

$$\frac{20}{(s^2+4)(s^2+2s+2)} = \frac{As+B}{s^2+4} + \frac{Ms+N}{s^2+2s+2}.$$

Multiplication by the common denominator gives

$$20 = (As + B)(s^2 + 2s + 2) + (Ms + N)(s^2 + 4).$$

We determine A, B, M, N. Equating the coefficients of each power of s on both sides gives the four equations

$$\begin{array}{llll} \text{(a)} \ [s^3]: & 0 = A + M & \text{(b)} \ [s^2]: & 0 = 2A + B + N \\ \text{(c)} \ [s]: & 0 = 2A + 2B + 4M & \text{(d)} \ [s^0]: & 20 = 2B + 4N. \end{array}$$

We can solve this, for instance, obtaining $M = -A$ from (a), then $A = B$ from (c), then $N = -3A$ from (b), and finally $A = -2$ from (d). Hence $A = -2$, $B = -2$, $M = 2$, $N = 6$, and the first fraction in (6) has the representation

$$\frac{-2s-2}{s^2+4} + \frac{2(s+1)+6-2}{(s+1)^2+1}. \qquad \text{Inverse transform:} \quad -2\cos 2t - \sin 2t + e^{-t}(2\cos t + 4 \sin t). \tag{8}$$

The sum of this inverse and (7) is the solution of the problem for $0 < t < \pi$, namely (the sines cancel),

$$y(t) = 3e^{-t}\cos t - 2\cos 2t - \sin 2t \qquad \text{if } 0 < t < \pi. \tag{9}$$

In the second fraction in (6), taken with the minus sign, we have the factor $e^{-\pi s}$, so that from (8) and the second shifting theorem (Sec. 6.3) we get the inverse transform of this fraction for $t > 0$ in the form

$$\begin{aligned} &+2\cos(2t - 2\pi) + \sin(2t - 2\pi) - e^{-(t-\pi)}[2\cos(t-\pi) + 4\sin(t-\pi)] \\ &= 2\cos 2t + \sin 2t + e^{-(t-\pi)}(2\cos t + 4\sin t). \end{aligned}$$

The sum of this and (9) is the solution for $t > \pi$,

$$y(t) = e^{-t}[(3 + 2e^{\pi})\cos t + 4e^{\pi}\sin t] \qquad \text{if } t > \pi. \tag{10}$$

Figure 136 shows (9) (for $0 < t < \pi$) and (10) (for $t > \pi$), a beginning vibration, which goes to zero rapidly because of the damping and the absence of a driving force after $t = \pi$. ■

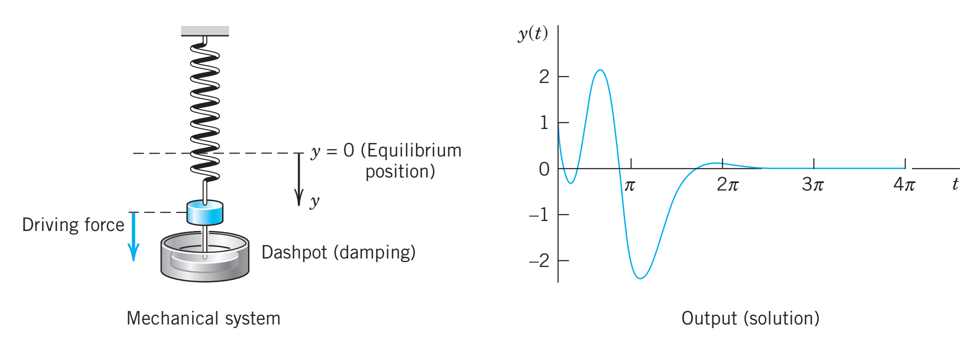


Fig. 136. Example 4

The case of repeated complex factors $[(s - a)(s - \bar{a})]^2$, which is important in connection with resonance, will be handled by "convolution" in the next section.

PROBLEM SET 6.4

1. CAS PROJECT. Effect of Damping. Consider a vibrating system of your choice modeled by

$$y'' + cy' + ky = \delta(t).$$

(a) Using graphs of the solution, describe the effect of continuously decreasing the damping to 0, keeping k constant.

(b) What happens if c is kept constant and k is continuously increased, starting from 0?

(c) Extend your results to a system with two δ-functions on the right, acting at different times.

2. CAS EXPERIMENT. Limit of a Rectangular Wave. Effects of Impulse.

(a) In Example 1 in the text, take a rectangular wave of area 1 from 1 to $1 + k$. Graph the responses for a sequence of values of k approaching zero, illustrating that for smaller and smaller k those curves approach the curve shown in Fig. 134. *Hint:* If your CAS gives no solution for the differential equation, involving k, take specific k's from the beginning.

(b) Experiment on the response of the ODE in Example 1 (or of another ODE of your choice) to an impulse $\delta(t - a)$ for various systematically chosen a (> 0); choose initial conditions $y(0) \neq 0$, $y'(0) = 0$. Also consider the solution if no impulse is applied. Is there a dependence of the response on a? On b if you choose $b\delta(t - a)$? Would $-\delta(t - \tilde{a})$ with $\tilde{a} > a$ annihilate the effect of $\delta(t - a)$? Can you think of other questions that one could consider experimentally by inspecting graphs?

3–12 EFFECT OF DELTA (IMPULSE) ON VIBRATING SYSTEMS

Find and graph or sketch the solution of the IVP. Show the details.

3. $y'' + 4y = \delta(t - \pi)$, $\quad y(0) = 8, y'(0) = 0$

4. $y'' + 16y = 4\delta(t - 3\pi), \quad y(0) = 2, y'(0) = 0$

5. $y'' + y = \delta(t - \pi) - \delta(t - 2\pi),$
$y(0) = 0, y'(0) = 1$

6. $y'' + 4y' + 5y = \delta(t - 1), \quad y(0) = 0, y'(0) = 3$

7. $4y'' + 24y' + 37y = 17e^{-t} + \delta(t - \frac{1}{2}),$
$y(0) = 1, y'(0) = 1$

8. $y'' + 3y' + 2y = 10(\sin t + \delta(t - 1)), \quad y(0) = 1,$
$y'(0) = -1$

9. $y'' + 4y' + 5y = [1 - u(t - 10)]e^t - e^{10}\delta(t - 10),$
$y(0) = 0, y'(0) = 1$

10. $y'' + 5y' + 6y = \delta(t - \frac{1}{2}\pi) + u(t - \pi)\cos t,$
$y(0) = 0, y'(0) = 0$

11. $y'' + 5y' + 6y = u(t - 1) + \delta(t - 2),$
$y(0) = 0, y'(0) = 1$

12. $y'' + 2y' + 5y = 25t - 100\delta(t - \pi), \quad y(0) = -2,$
$y'(0) = 5$

13. PROJECT. Heaviside Formulas. (a) Show that for a simple root a and fraction $A/(s - a)$ in $F(s)/G(s)$ we have the *Heaviside formula*

$$A = \lim_{s \to a} \frac{(s - a)F(s)}{G(s)}.$$

(b) Similarly, show that for a root a of order m and fractions in

$$\frac{F(s)}{G(s)} = \frac{A_m}{(s - a)^m} + \frac{A_{m-1}}{(s - a)^{m-1}} + \cdots$$
$$+ \frac{A_1}{s - a} + \text{further fractions}$$

we have the *Heaviside formulas* for the first coefficient

$$A_m = \lim_{s \to a} \frac{(s - a)^m F(s)}{G(s)}$$

and for the other coefficients

$$A_k = \frac{1}{(m - k)!} \lim_{s \to a} \frac{d^{m-k}}{ds^{m-k}} \left[\frac{(s - a)^m F(s)}{G(s)} \right],$$
$$k = 1, \cdots, m - 1.$$

14. TEAM PROJECT. Laplace Transform of Periodic Functions

(a) Theorem. *The Laplace transform of a piecewise continuous function $f(t)$ with period p is*

(11) $$\mathcal{L}(f) = \frac{1}{1 - e^{-ps}} \int_0^p e^{-st} f(t)\, dt \qquad (s > 0).$$

Prove this theorem. *Hint:* Write $\int_0^\infty = \int_0^p + \int_p^{2p} + \cdots$. Set $t = (n - 1)p$ in the nth integral. Take out $e^{-(n-1)p}$ from under the integral sign. Use the sum formula for the geometric series.

(b) Half-wave rectifier. Using (11), show that the half-wave rectification of $\sin \omega t$ in Fig. 137 has the Laplace transform

$$\mathcal{L}(f) = \frac{\omega(1 + e^{-\pi s/\omega})}{(s^2 + \omega^2)(1 - e^{-2\pi s/\omega})}$$
$$= \frac{\omega}{(s^2 + \omega^2)(1 - e^{-\pi s/\omega})}.$$

(A *half-wave rectifier* clips the negative portions of the curve. A *full-wave rectifier* converts them to positive; see Fig. 138.)

(c) Full-wave rectifier. Show that the Laplace transform of the full-wave rectification of $\sin \omega t$ is

$$\frac{\omega}{s^2 + \omega^2} \coth \frac{\pi s}{2\omega}.$$

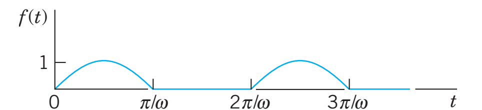


Fig. 137. Half-wave rectification

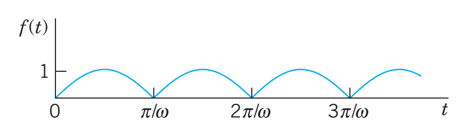


Fig. 138. Full-wave rectification

(d) Saw-tooth wave. Find the Laplace transform of the saw-tooth wave in Fig. 139.

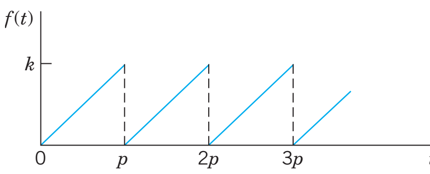


Fig. 139. Saw-tooth wave

15. Staircase function. Find the Laplace transform of the staircase function in Fig. 140 by noting that it is the difference of kt/p and the function in 14(d).

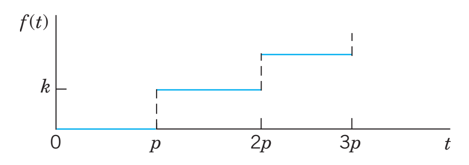


Fig. 140. Staircase function

6.5 Convolution. Integral Equations

Convolution has to do with the multiplication of transforms. The situation is as follows. *Addition* of transforms provides no problem; we know that $\mathcal{L}(f+g) = \mathcal{L}(f) + \mathcal{L}(g)$. Now **multiplication of transforms** occurs frequently in connection with ODEs, integral equations, and elsewhere. Then we usually know $\mathcal{L}(f)$ and $\mathcal{L}(g)$ and would like to know the function whose transform is the product $\mathcal{L}(f)\mathcal{L}(g)$. We might perhaps guess that it is fg, but this is false. *The transform of a product is generally different from the product of the transforms of the factors,*

$$\mathcal{L}(fg) \neq \mathcal{L}(f)\mathcal{L}(g) \qquad \text{in general.}$$

To see this take $f = e^t$ and $g = 1$. Then $fg = e^t$, $\mathcal{L}(fg) = 1/(s-1)$, but $\mathcal{L}(f) = 1/(s-1)$ and $\mathcal{L}(1) = 1/s$ give $\mathcal{L}(f)\mathcal{L}(g) = 1/(s^2 - s)$.

According to the next theorem, the correct answer is that $\mathcal{L}(f)\mathcal{L}(g)$ is the transform of the **convolution** of f and g, denoted by the standard notation $f * g$ and defined by the integral

$$h(t) = (f * g)(t) = \int_0^t f(\tau)g(t-\tau)\,d\tau. \tag{1}$$

THEOREM 1

Convolution Theorem

If two functions f and g satisfy the assumption in the existence theorem in Sec. 6.1, *so that their transforms F and G exist, the product $H = FG$ is the transform of h given by* (1). (Proof after Example 2.)

EXAMPLE 1

Convolution

Let $H(s) = 1/[(s-a)s]$. Find $h(t)$.

Solution. $1/(s-a)$ has the inverse $f(t) = e^{at}$, and $1/s$ has the inverse $g(t) = 1$. With $f(\tau) = e^{a\tau}$ and $g(t-\tau) \equiv 1$ we thus obtain from (1) the answer

$$h(t) = e^{at} * 1 = \int_0^t e^{a\tau} \cdot 1\,d\tau = \frac{1}{a}(e^{at} - 1).$$

To check, calculate

$$H(s) = \mathcal{L}(h)(s) = \frac{1}{a}\left(\frac{1}{s-a} - \frac{1}{s}\right) = \frac{1}{a} \cdot \frac{a}{s^2 - as} = \frac{1}{s-a} \cdot \frac{1}{s} = \mathcal{L}(e^{at})\mathcal{L}(1). \qquad \blacksquare$$

EXAMPLE 2

Convolution

Let $H(s) = 1/(s^2 + \omega^2)^2$. Find $h(t)$.

Solution. The inverse of $1/(s^2 + \omega^2)$ is $(\sin \omega t)/\omega$. Hence from (1) and the first formula in (11) in App. 3.1 we obtain

$$\begin{aligned} h(t) = \frac{\sin \omega t}{\omega} * \frac{\sin \omega t}{\omega} &= \frac{1}{\omega^2}\int_0^t \sin \omega\tau \sin \omega(t-\tau)\,d\tau \\ &= \frac{1}{2\omega^2}\int_0^t [-\cos \omega t + \cos(2\omega\tau - \omega t)]\,d\tau \end{aligned}$$

$$= \frac{1}{2\omega^2}\left[-\tau \cos \omega t + \frac{\sin \omega\tau}{\omega}\right]_{\tau=0}^{t}$$

$$= \frac{1}{2\omega^2}\left[-t \cos \omega t + \frac{\sin \omega t}{\omega}\right]$$

in agreement with formula 21 in the table in Sec. 6.9. ■

PROOF We prove the Convolution Theorem 1. CAUTION! Note which ones are the variables of integration! We can denote them as we want, for instance, by τ and p, and write

$$F(s) = \int_0^\infty e^{-s\tau} f(\tau)\, d\tau \qquad \text{and} \qquad G(s) = \int_0^\infty e^{-sp} g(p)\, dp.$$

We now set $t = p + \tau$, where τ is at first constant. Then $p = t - \tau$, and t varies from τ to ∞. Thus

$$G(s) = \int_\tau^\infty e^{-s(t-\tau)} g(t - \tau)\, dt = e^{s\tau} \int_\tau^\infty e^{-st} g(t - \tau)\, dt.$$

τ in F and t in G vary independently. Hence we can insert the G-integral into the F-integral. Cancellation of $e^{-s\tau}$ and $e^{s\tau}$ then gives

$$F(s)G(s) = \int_0^\infty e^{-s\tau} f(\tau) e^{s\tau} \int_\tau^\infty e^{-st} g(t - \tau)\, dt\, d\tau = \int_0^\infty f(\tau) \int_\tau^\infty e^{-st} g(t - \tau)\, dt\, d\tau.$$

Here we integrate for fixed τ over t from τ to ∞ and then over τ from 0 to ∞. This is the blue region in Fig. 141. Under the assumption on f and g the order of integration can be reversed (see Ref. [A5] for a proof using uniform convergence). We then integrate first over τ from 0 to t and then over t from 0 to ∞, that is,

$$F(s)G(s) = \int_0^\infty e^{-st} \int_0^t f(\tau) g(t - \tau)\, d\tau\, dt = \int_0^\infty e^{-st} h(t)\, dt = \mathcal{L}(h) = H(s).$$

This completes the proof. ■

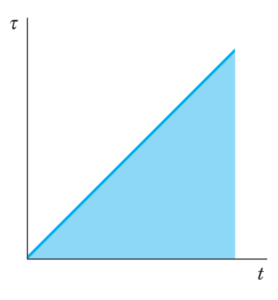


Fig. 141. Region of integration in the $t\tau$-plane in the proof of Theorem 1

From the definition it follows almost immediately that convolution has the properties

$$f * g = g * f \qquad \text{(commutative law)}$$

$$f * (g_1 + g_2) = f * g_1 + f * g_2 \qquad \text{(distributive law)}$$

$$(f * g) * v = f * (g * v) \qquad \text{(associative law)}$$

$$f * 0 = 0 * f = 0$$

similar to those of the multiplication of numbers. However, there are differences of which you should be aware.

EXAMPLE 3 **Unusual Properties of Convolution**

$f * 1 \neq f$ in general. For instance,

$$t * 1 = \int_0^t \tau \cdot 1\, d\tau = \frac{1}{2} t^2 \neq t.$$

$(f * f)(t) \geqq 0$ may not hold. For instance, Example 2 with $\omega = 1$ gives

$$\sin t * \sin t = -\tfrac{1}{2} t \cos t + \tfrac{1}{2} \sin t \qquad \text{(Fig. 142).}$$ ■

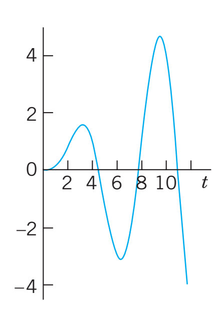


Fig. 142. Example 3

We shall now take up the case of a complex double root (left aside in the last section in connection with partial fractions) and find the solution (the inverse transform) directly by convolution.

EXAMPLE 4 **Repeated Complex Factors. Resonance**

In an undamped mass–spring system, resonance occurs if the frequency of the driving force equals the natural frequency of the system. Then the model is (see Sec. 2.8)

$$y'' + \omega_0^2 y = K \sin \omega_0 t$$

where $\omega_0^2 = k/m$, k is the spring constant, and m is the mass of the body attached to the spring. We assume $y(0) = 0$ and $y'(0) = 0$, for simplicity. Then the subsidiary equation is

$$s^2 Y + \omega_0^2 Y = \frac{K\omega_0}{s^2 + \omega_0^2}. \qquad \text{Its solution is} \qquad Y = \frac{K\omega_0}{(s^2 + \omega_0^2)^2}.$$

This is a transform as in Example 2 with $\omega = \omega_0$ and multiplied by $K\omega_0$. Hence from Example 2 we can see directly that the solution of our problem is

$$y(t) = \frac{K\omega_0}{2\omega_0^2}\left(-t\cos\omega_0 t + \frac{\sin\omega_0 t}{\omega_0}\right) = \frac{K}{2\omega_0^2}(-\omega_0 t\cos\omega_0 t + \sin\omega_0 t).$$

We see that the first term grows without bound. Clearly, in the case of resonance such a term must occur. (See also a similar kind of solution in Fig. 55 in Sec. 2.8.) ■

Application to Nonhomogeneous Linear ODEs

Nonhomogeneous linear ODEs can now be solved by a general method based on convolution by which the solution is obtained in the form of an integral. To see this, recall from Sec. 6.2 that the subsidiary equation of the ODE

$$y'' + ay' + by = r(t) \qquad (a,\ b \text{ constant}) \tag{2}$$

has the solution [(7) in Sec. 6.2]

$$Y(s) = [(s+a)y(0) + y'(0)]Q(s) + R(s)Q(s)$$

with $R(s) = \mathcal{L}(r)$ and $Q(s) = 1/(s^2 + as + b)$ the transfer function. Inversion of the first term $[\cdots]$ provides no difficulty; depending on whether $\frac{1}{4}a^2 - b$ is positive, zero, or negative, its inverse will be a linear combination of two exponential functions, or of the form $(c_1 + c_2 t)e^{-at/2}$, or a damped oscillation, respectively. The interesting term is $R(s)Q(s)$ because $r(t)$ can have various forms of practical importance, as we shall see. If $y(0) = 0$ and $y'(0) = 0$, then $Y = RQ$, and the convolution theorem gives the solution

$$y(t) = \int_0^t q(t-\tau)r(\tau)\,d\tau. \tag{3}$$

EXAMPLE 5 **Response of a Damped Vibrating System to a Single Square Wave**

Using convolution, determine the response of the damped mass–spring system modeled by

$$y'' + 3y' + 2y = r(t), \qquad r(t) = 1 \text{ if } 1 < t < 2 \text{ and 0 otherwise}, \qquad y(0) = y'(0) = 0.$$

This system with an **input** (a driving force) ***that acts for some time only*** (Fig. 143) has been solved by partial fraction reduction in Sec. 6.4 (Example 1).

Solution by Convolution. The transfer function and its inverse are

$$Q(s) = \frac{1}{s^2 + 3s + 2} = \frac{1}{(s+1)(s+2)} = \frac{1}{s+1} - \frac{1}{s+2}, \qquad \text{hence} \qquad q(t) = e^{-t} - e^{-2t}.$$

Hence the convolution integral (3) is (except for the limits of integration)

$$y(t) = \int q(t-\tau)\cdot 1\,d\tau = \int \left[e^{-(t-\tau)} - e^{-2(t-\tau)}\right]d\tau = e^{-(t-\tau)} - \frac{1}{2}e^{-2(t-\tau)}.$$

Now comes an important point in handling convolution. $r(\tau) = 1$ if $1 < \tau < 2$ only. Hence if $t < 1$, the integral is zero. If $1 < t < 2$, we have to integrate from $\tau = 1$ (not 0) to t. This gives (with the first two terms from the upper limit)

$$y(t) = e^{-0} - \tfrac{1}{2}e^{-0} - (e^{-(t-1)} - \tfrac{1}{2}e^{-2(t-1)}) = \tfrac{1}{2} - e^{-(t-1)} + \tfrac{1}{2}e^{-2(t-1)}.$$

If $t > 2$, we have to integrate from $\tau = 1$ to 2 (not to t). This gives

$$y(t) = e^{-(t-2)} - \tfrac{1}{2}e^{-2(t-2)} - (e^{-(t-1)} - \tfrac{1}{2}e^{-2(t-1)}).$$

Figure 143 shows the input (the square wave) and the interesting output, which is zero from 0 to 1, then increases, reaches a maximum (near 2.6) after the input has become zero (why?), and finally decreases to zero in a monotone fashion.

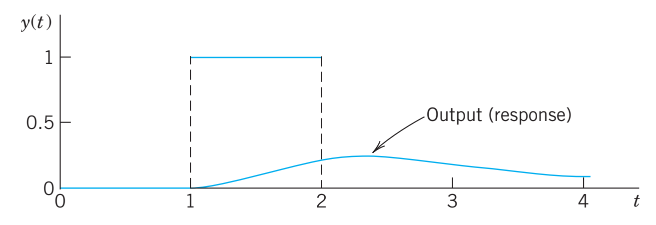


Fig. 143. Square wave and response in Example 5

Integral Equations

Convolution also helps in solving certain **integral equations**, that is, equations in which the unknown function $y(t)$ appears in an integral (and perhaps also outside of it). This concerns equations with an integral of the form of a convolution. Hence these are special and it suffices to explain the idea in terms of two examples and add a few problems in the problem set.

EXAMPLE 6 A Volterra Integral Equation of the Second Kind

Solve the Volterra integral equation of the second kind[3]

$$y(t) - \int_0^t y(\tau)\sin(t-\tau)\,d\tau = t.$$

Solution. From (1) we see that the given equation can be written as a convolution, $y - y * \sin t = t$. Writing $Y = \mathcal{L}(y)$ and applying the convolution theorem, we obtain

$$Y(s) - Y(s)\frac{1}{s^2+1} = Y(s)\frac{s^2}{s^2+1} = \frac{1}{s^2}.$$

The solution is

$$Y(s) = \frac{s^2+1}{s^4} = \frac{1}{s^2} + \frac{1}{s^4} \qquad \text{and gives the answer} \qquad y(t) = t + \frac{t^3}{6}.$$

Check the result by a CAS or by substitution and repeated integration by parts (which will need patience). ■

EXAMPLE 7 Another Volterra Integral Equation of the Second Kind

Solve the Volterra integral equation

$$y(t) - \int_0^t (1+\tau)\,y(t-\tau)\,d\tau = 1 - \sinh t.$$

[3]If the upper limit of integration is *variable*, the equation is named after the Italian mathematician VITO VOLTERRA (1860–1940), and if that limit is *constant*, the equation is named after the Swedish mathematician ERIK IVAR FREDHOLM (1866–1927). "Of the second kind (first kind)" indicates that y occurs (does not occur) outside of the integral.

Solution. By (1) we can write $y - (1 + t) * y = 1 - \sinh t$. Writing $Y = \mathcal{L}(y)$, we obtain by using the convolution theorem and then taking common denominators

$$Y(s)\left[1 - \left(\frac{1}{s} + \frac{1}{s^2}\right)\right] = \frac{1}{s} - \frac{1}{s^2 - 1}, \qquad \text{hence} \qquad Y(s) \cdot \frac{s^2 - s - 1}{s^2} = \frac{s^2 - 1 - s}{s(s^2 - 1)}.$$

$(s^2 - s - 1)/s$ cancels on both sides, so that solving for Y simply gives

$$Y(s) = \frac{s}{s^2 - 1} \qquad \text{and the solution is} \qquad y(t) = \cosh t.$$

PROBLEM SET 6.5

1–7 CONVOLUTIONS BY INTEGRATION

Find:

1. $1 * 1$
2. $1 * \sin \omega t$
3. $e^t * e^{-t}$
4. $(\cos \omega t) * (\cos \omega t)$
5. $(\sin \omega t) * (\cos \omega t)$
6. $e^{at} * e^{bt}\ (a \neq b)$
7. $t * e^t$

8–14 INTEGRAL EQUATIONS

Solve by the Laplace transform, showing the details:

8. $y(t) + 4 \int_0^t y(\tau)(t - \tau)\, d\tau = 2t$
9. $y(t) - \int_0^t y(\tau)\, d\tau = 1$
10. $y(t) - \int_0^t y(\tau) \sin 2(t - \tau)\, d\tau = \sin 2t$
11. $y(t) + \int_0^t (t - \tau) y(\tau)\, d\tau = 1$
12. $y(t) + \int_0^t y(\tau) \cosh (t - \tau)\, d\tau = t + e^t$
13. $y(t) + 2e^t \int_0^t y(\tau) e^{-\tau}\, d\tau = te^t$
14. $y(t) - \int_0^t y(\tau)(t - \tau)\, d\tau = 2 - \frac{1}{2}t^2$

15. **CAS EXPERIMENT. Variation of a Parameter.**
(a) Replace 2 in Prob. 13 by a parameter k and investigate graphically how the solution curve changes if you vary k, in particular near $k = -2$.
(b) Make similar experiments with an integral equation of your choice whose solution is oscillating.

16. **TEAM PROJECT. Properties of Convolution.** Prove:
(a) Commutativity, $f * g = g * f$
(b) Associativity, $(f * g) * v = f * (g * v)$
(c) Distributivity, $f * (g_1 + g_2) = f * g_1 + f * g_2$
(d) **Dirac's delta.** Derive the sifting formula (4) in Sec. 6.4 by using f_k with $a = 0$ [(1), Sec. 6.4] and applying the mean value theorem for integrals.
(e) **Unspecified driving force.** Show that forced vibrations governed by
$$y'' + \omega^2 y = r(t), \quad y(0) = K_1, \quad y'(0) = K_2$$
with $\omega \neq 0$ and an unspecified driving force $r(t)$ can be written in convolution form,
$$y = \frac{1}{\omega} \sin \omega t * r(t) + K_1 \cos \omega t + \frac{K_2}{\omega} \sin \omega t.$$

17–26 INVERSE TRANSFORMS BY CONVOLUTION

Showing details, find $f(t)$ if $\mathcal{L}(f)$ equals:

17. $\dfrac{5.5}{(s + 1.5)(s - 4)}$
18. $\dfrac{1}{(s - a)^2}$
19. $\dfrac{2\pi s}{(s^2 + \pi^2)^2}$
20. $\dfrac{9}{s(s + 3)}$
21. $\dfrac{\omega}{s^2(s^2 + \omega^2)}$
22. $\dfrac{e^{-as}}{s(s - 2)}$
23. $\dfrac{40.5}{s(s^2 - 9)}$
24. $\dfrac{240}{(s^2 + 1)(s^2 + 25)}$
25. $\dfrac{18s}{(s^2 + 36)^2}$
26. **Partial Fractions.** Solve Probs. 17, 21, and 23 by partial fraction reduction.

6.6 Differentiation and Integration of Transforms. ODEs with Variable Coefficients

The variety of methods for obtaining transforms and inverse transforms and their application in solving ODEs is surprisingly large. We have seen that they include direct integration, the use of linearity (Sec. 6.1), shifting (Secs. 6.1, 6.3), convolution (Sec. 6.5), and differentiation and integration of functions $f(t)$ (Sec. 6.2). In this section, we shall consider operations of somewhat lesser importance. They are the differentiation and integration of transforms $F(s)$ and corresponding operations for functions $f(t)$. We show how they are applied to ODEs with variable coefficients.

Differentiation of Transforms

It can be shown that, if a function $f(t)$ satisfies the conditions of the existence theorem in Sec. 6.1, then the derivative $F'(s) = dF/ds$ of the transform $F(s) = \mathcal{L}(f)$ can be obtained by differentiating $F(s)$ under the integral sign with respect to s (proof in Ref. [GenRef4] listed in App. 1). Thus, if

$$F(s) = \int_0^\infty e^{-st} f(t)\, dt, \qquad \text{then} \qquad F'(s) = -\int_0^\infty e^{-st} t f(t)\, dt.$$

Consequently, if $\mathcal{L}(f) = F(s)$, then

$$\textbf{(1)} \qquad \mathcal{L}\{tf(t)\} = -F'(s), \qquad \text{hence} \qquad \mathcal{L}^{-1}\{F'(s)\} = -tf(t)$$

where the second formula is obtained by applying $\mathcal{L}^{-1}$ on both sides of the first formula. In this way, *differentiation of the transform of a function corresponds to the multiplication of the function by* $-t$.

EXAMPLE 1 Differentiation of Transforms. Formulas 21–23 in Sec. 6.9

We shall derive the following three formulas.

	$\mathcal{L}(f)$	$f(t)$
(2)	$\dfrac{1}{(s^2+\beta^2)^2}$	$\dfrac{1}{2\beta^3}(\sin \beta t - \beta t \cos \beta t)$
(3)	$\dfrac{s}{(s^2+\beta^2)^2}$	$\dfrac{1}{2\beta} \sin \beta t$
(4)	$\dfrac{s^2}{(s^2+\beta^2)^2}$	$\dfrac{1}{2\beta}(\sin \beta t + \beta t \cos \beta t)$

Solution. From (1) and formula 8 (with $\omega = \beta$) in Table 6.1 of Sec. 6.1 we obtain by differentiation (CAUTION! Chain rule!)

$$\mathcal{L}(t \sin \beta t) = \frac{2\beta s}{(s^2+\beta^2)^2}.$$

Dividing by 2β and using the linearity of $\mathcal{L}$, we obtain (3).

Formulas (2) and (4) are obtained as follows. From (1) and formula 7 (with $\omega = \beta$) in Table 6.1 we find

$$\text{(5)} \qquad \mathcal{L}(t \cos \beta t) = -\frac{(s^2 + \beta^2) - 2s^2}{(s^2 + \beta^2)^2} = \frac{s^2 - \beta^2}{(s^2 + \beta^2)^2}.$$

From this and formula 8 (with $\omega = \beta$) in Table 6.1 we have

$$\mathcal{L}\left(t \cos \beta t \pm \frac{1}{\beta} \sin \beta t\right) = \frac{s^2 - \beta^2}{(s^2 + \beta^2)^2} \pm \frac{1}{s^2 + \beta^2}.$$

On the right we now take the common denominator. Then we see that for the plus sign the numerator becomes $s^2 - \beta^2 + s^2 + \beta^2 = 2s^2$, so that (4) follows by division by 2. Similarly, for the minus sign the numerator takes the form $s^2 - \beta^2 - s^2 - \beta^2 = -2\beta^2$, and we obtain (2). This agrees with Example 2 in Sec. 6.5. ■

Integration of Transforms

Similarly, if $f(t)$ satisfies the conditions of the existence theorem in Sec. 6.1 and the limit of $f(t)/t$, as t approaches 0 from the right, exists, then for $s > k$,

$$\textbf{(6)} \qquad \mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\tilde{s})\, d\tilde{s} \qquad \text{hence} \qquad \mathcal{L}^{-1}\left\{\int_s^\infty F(\tilde{s})\, d\tilde{s}\right\} = \frac{f(t)}{t}.$$

In this way, *integration of the transform of a function* $f(t)$ *corresponds to the division of* $f(t)$ *by* t.

We indicate how (6) is obtained. From the definition it follows that

$$\int_s^\infty F(\tilde{s})\, d\tilde{s} = \int_s^\infty \left[\int_0^\infty e^{-\tilde{s}t} f(t)\, dt\right] d\tilde{s},$$

and it can be shown (see Ref. [GenRef4] in App. 1) that under the above assumptions we may reverse the order of integration, that is,

$$\int_s^\infty F(\tilde{s})\, d\tilde{s} = \int_0^\infty \left[\int_s^\infty e^{-\tilde{s}t} f(t)\, d\tilde{s}\right] dt = \int_0^\infty f(t) \left[\int_s^\infty e^{-\tilde{s}t}\, d\tilde{s}\right] dt.$$

Integration of $e^{-\tilde{s}t}$ with respect to $\tilde{s}$ gives $e^{-\tilde{s}t}/(-t)$. Here the integral over $\tilde{s}$ on the right equals e^{-st}/t. Therefore,

$$\int_s^\infty F(\tilde{s})\, d\tilde{s} = \int_0^\infty e^{-st} \frac{f(t)}{t}\, dt = \mathcal{L}\left\{\frac{f(t)}{t}\right\} \qquad (s > k).$$ ■

EXAMPLE 2 Differentiation and Integration of Transforms

Find the inverse transform of $\ln\left(1 + \frac{\omega^2}{s^2}\right) = \ln \frac{s^2 + \omega^2}{s^2}$.

Solution. Denote the given transform by $F(s)$. Its derivative is

$$F'(s) = \frac{d}{ds}(\ln (s^2 + \omega^2) - \ln s^2) = \frac{2s}{s^2 + \omega^2} - \frac{2s}{s^2}.$$

Taking the inverse transform and using (1), we obtain

$$\mathcal{L}^{-}\{F'(s)\} = \mathcal{L}^{-1}\left\{\frac{2s}{s^2+\omega^2} - \frac{2}{s}\right\} = 2\cos\omega t - 2 = -tf(t).$$

Hence the inverse $f(t)$ of $F(s)$ is $f(t) = 2(1 - \cos\omega t)/t$. This agrees with formula 42 in Sec. 6.9.
Alternatively, if we let

$$G(s) = \frac{2s}{s^2+\omega^2} - \frac{2}{s}, \qquad \text{then} \qquad g(t) = \mathcal{L}^{-1}(G) - 2(\cos\omega t - 1).$$

From this and (6) we get, in agreement with the answer just obtained,

$$\mathcal{L}^{-1}\left\{\ln\frac{s^2+\omega^2}{s^2}\right\} = \mathcal{L}^{-1}\left\{\int_s^{\infty} G(s)\,ds\right\} = -\frac{g(t)}{t} = \frac{2}{t}(1-\cos\omega t),$$

the minus occurring since s is the lower limit of integration.
In a similar way we obtain formula 43 in Sec. 6.9,

$$\mathcal{L}^{-1}\left\{\ln\left(1-\frac{a^2}{s^2}\right)\right\} = \frac{2}{t}(1-\cosh at).$$ ■

Special Linear ODEs with Variable Coefficients

Formula (1) can be used to solve certain ODEs with variable coefficients. The idea is this. Let $\mathcal{L}(y) = Y$. Then $\mathcal{L}(y') = sY - y(0)$ (see Sec. 6.2). Hence by (1),

$$\text{(7)} \qquad \mathcal{L}(ty') = -\frac{d}{ds}[sY - y(0)] = -Y - s\frac{dY}{ds}.$$

Similarly, $\mathcal{L}(y'') = s^2Y - sy(0) - y'(0)$ and by (1)

$$\text{(8)} \qquad \mathcal{L}(ty'') = -\frac{d}{ds}[s^2Y - sy(0) - y'(0)] = -2sY - s^2\frac{dY}{ds} + y(0).$$

Hence if an ODE has coefficients such as $at + b$, the subsidiary equation is a first-order ODE for Y, which is sometimes simpler than the given second-order ODE. But if the latter has coefficients $at^2 + bt + c$, then two applications of (1) would give a second-order ODE for Y, and this shows that the present method works well only for rather special ODEs with variable coefficients. An important ODE for which the method is advantageous is the following.

EXAMPLE 3 Laguerre's Equation. Laguerre Polynomials

Laguerre's ODE is

$$\textbf{(9)} \qquad ty'' + (1-t)y' + ny = 0.$$

We determine a solution of (9) with $n = 0, 1, 2, \cdots$. From (7)–(9) we get the subsidiary equation

$$\left[-2sY - s^2\frac{dY}{ds} + y(0)\right] + sY - y(0) - \left(-Y - s\frac{dY}{ds}\right) + nY = 0.$$

Simplification gives

$$(s - s^2)\frac{dY}{ds} + (n + 1 - s)Y = 0.$$

Separating variables, using partial fractions, integrating (with the constant of integration taken to be zero), and taking exponentials, we get

$$(10^*) \qquad \frac{dY}{Y} = -\frac{n + 1 - s}{s - s^2}\,ds = \left(\frac{n}{s - 1} - \frac{n + 1}{s}\right)ds \qquad \text{and} \qquad Y = \frac{(s - 1)^n}{s^{n+1}}.$$

We write $l_n = \mathcal{L}^{-1}(Y)$ and prove **Rodrigues's formula**

$$(10) \qquad l_0 = 1, \qquad l_n(t) = \frac{e^t}{n!}\frac{d^n}{dt^n}(t^n e^{-t}), \qquad n = 1, 2, \cdots.$$

These are polynomials because the exponential terms cancel if we perform the indicated differentiations. They are called **Laguerre polynomials** and are usually denoted by L_n (see Problem Set 5.7, but we continue to reserve capital letters for transforms). We prove (10). By Table 6.1 and the first shifting theorem (s-shifting),

$$\mathcal{L}(t^n e^{-t}) = \frac{n!}{(s + 1)^{n+1}}, \qquad \text{hence by (3) in Sec. 6.2} \qquad \mathcal{L}\left\{\frac{d^n}{dt^n}(t^n e^{-t})\right\} = \frac{n!s^n}{(s + 1)^{n+1}}$$

because the derivatives up to the order $n - 1$ are zero at 0. Now make another shift and divide by $n!$ to get [see (10) and then (10*)]

$$\mathcal{L}(l_n) = \frac{(s - 1)^n}{s^{n+1}} = Y. \qquad \blacksquare$$

PROBLEM SET 6.6

1. **REVIEW REPORT. Differentiation and Integration of Functions and Transforms.** Make a draft of these four operations from memory. Then compare your draft with the text and write a 2- to 3-page report on these operations and their significance in applications.

2–11 TRANSFORMS BY DIFFERENTIATION

Showing the details of your work, find $\mathcal{L}(f)$ if $f(t)$ equals:

2. $3t \sinh 4t$
3. $\frac{1}{2}te^{-3t}$
4. $te^{-t} \cos t$
5. $t \cos \omega t$
6. $t^2 \sin 3t$
7. $t^2 \cosh 2t$
8. $te^{-kt} \sin t$
9. $\frac{1}{2}t^2 \sin \pi t$
10. $t^n e^{kt}$
11. $4t \cos \frac{1}{2}\pi t$
12. **CAS PROJECT. Laguerre Polynomials. (a)** Write a CAS program for finding $l_n(t)$ in explicit form from (10). Apply it to calculate $l_0, \cdots, l_{10}$. Verify that $l_0, \cdots, l_{10}$ satisfy Laguerre's differential equation (9).

(b) Show that

$$l_n(t) = \sum_{m=0}^{n} \frac{(-1)^m}{m!}\binom{n}{m} t^m$$

and calculate $l_0, \cdots, l_{10}$ from this formula.

(c) Calculate $l_0, \cdots, l_{10}$ recursively from $l_0 = 1$, $l_1 = 1 - t$ by

$$(n + 1)l_{n+1} = (2n + 1 - t)l_n - nl_{n-1}.$$

(d) A **generating function** (definition in Problem Set 5.2) for the Laguerre polynomials is

$$\sum_{n=0}^{\infty} l_n(t)x^n = (1 - x)^{-1}e^{tx/(x-1)}.$$

Obtain $l_0, \cdots, l_{10}$ from the corresponding partial sum of this power series in x and compare the l_n with those in (a), (b), or (c).

13. **CAS EXPERIMENT. Laguerre Polynomials.** Experiment with the graphs of $l_0, \cdots, l_{10}$, finding out empirically how the first maximum, first minimum, $\cdots$ is moving with respect to its location as a function of n. Write a short report on this.

14–20 INVERSE TRANSFORMS

Using differentiation, integration, s-shifting, or convolution, and showing the details, find $f(t)$ if $\mathcal{L}(f)$ equals:

14. $\dfrac{s}{(s^2 + 16)^2}$

15. $\dfrac{s}{(s^2 - 9)^2}$

16. $\dfrac{2s + 6}{(s^2 + 6s + 10)^2}$

17. $\ln \dfrac{s}{s - 1}$

18. $\operatorname{arccot} \dfrac{s}{\pi}$

19. $\ln \dfrac{s^2 + 1}{(s - 1)^2}$

20. $\ln \dfrac{s + a}{s + b}$

6.7 Systems of ODEs

The Laplace transform method may also be used for solving systems of ODEs, as we shall explain in terms of typical applications. We consider a first-order linear system with constant coefficients (as discussed in Sec. 4.1)

$$\text{(1)} \qquad \begin{aligned} y_1' &= a_{11}y_1 + a_{12}y_2 + g_1(t) \\ y_2' &= a_{21}y_1 + a_{22}y_2 + g_2(t). \end{aligned}$$

Writing $Y_1 = \mathcal{L}(y_1)$, $Y_2 = \mathcal{L}(y_2)$, $G_1 = \mathcal{L}(g_1)$, $G_2 = \mathcal{L}(g_2)$, we obtain from (1) in Sec. 6.2 the subsidiary system

$$\begin{aligned} sY_1 - y_1(0) &= a_{11}Y_1 + a_{12}Y_2 + G_1(s) \\ sY_2 - y_2(0) &= a_{21}Y_1 + a_{22}Y_2 + G_2(s). \end{aligned}$$

By collecting the Y_1- and Y_2-terms we have

$$\textbf{(2)} \qquad \begin{aligned} (a_{11} - s)Y_1 + \quad a_{12}Y_2 \quad &= -y_1(0) - G_1(s) \\ a_{21}Y_1 \quad + (a_{22} - s)Y_2 &= -y_2(0) - G_2(s). \end{aligned}$$

By solving this system algebraically for $Y_1(s), Y_2(s)$ and taking the inverse transform we obtain the solution $y_1 = \mathcal{L}^{-1}(Y_1)$, $y_2 = \mathcal{L}^{-1}(Y_2)$ of the given system (1).

Note that (1) and (2) may be written in vector form (and similarly for the systems in the examples); thus, setting $\mathbf{y} = [y_1 \quad y_2]^\mathsf{T}$, $\mathbf{A} = [a_{jk}]$, $\mathbf{g} = [g_1 \quad g_2]^\mathsf{T}$, $\mathbf{Y} = [Y_1 \quad Y_2]^\mathsf{T}$, $\mathbf{G} = [G_1 \quad G_2]^\mathsf{T}$ we have

$$\mathbf{y}' = \mathbf{A}\mathbf{y} + \mathbf{g} \qquad \text{and} \qquad (\mathbf{A} - s\mathbf{I})\mathbf{Y} = -\mathbf{y}(0) - \mathbf{G}.$$

EXAMPLE 1 **Mixing Problem Involving Two Tanks**

Tank T_1 in Fig. 144 initially contains 100 gal of pure water. Tank T_2 initially contains 100 gal of water in which 150 lb of salt are dissolved. The inflow into T_1 is 2 gal/min from T_2 and 6 gal/min containing 6 lb of salt from the outside. The inflow into T_2 is 8 gal/min from T_1. The outflow from T_2 is $2 + 6 = 8$ gal/min, as shown in the figure. The mixtures are kept uniform by stirring. Find and plot the salt contents $y_1(t)$ and $y_2(t)$ in T_1 and T_2, respectively.

Solution. The model is obtained in the form of two equations

$$\text{Time rate of change} = \text{Inflow/min} - \text{Outflow/min}$$

for the two tanks (see Sec. 4.1). Thus,

$$y_1' = -\tfrac{8}{100}y_1 + \tfrac{2}{100}y_2 + 6. \qquad y_2' = \tfrac{8}{100}y_1 - \tfrac{8}{100}y_2.$$

The initial conditions are $y_1(0) = 0$, $y_2(0) = 150$. From this we see that the subsidiary system (2) is

$$\begin{aligned} (-0.08 - s)Y_1 + \quad 0.02Y_2 \quad &= -\frac{6}{s} \\ 0.08Y_1 \quad + (-0.08 - s)Y_2 &= -150. \end{aligned}$$

We solve this algebraically for Y_1 and Y_2 by elimination (or by Cramer's rule in Sec. 7.7), and we write the solutions in terms of partial fractions,

$$\begin{aligned} Y_1 &= \frac{9s + 0.48}{s(s + 0.12)(s + 0.04)} = \frac{100}{s} - \frac{62.5}{s + 0.12} - \frac{37.5}{s + 0.04} \\ Y_2 &= \frac{150s^2 + 12s + 0.48}{s(s + 0.12)(s + 0.04)} = \frac{100}{s} + \frac{125}{s + 0.12} - \frac{75}{s + 0.04}. \end{aligned}$$

By taking the inverse transform we arrive at the solution

$$\begin{aligned} y_1 &= 100 - 62.5e^{-0.12t} - 37.5e^{-0.04t} \\ y_2 &= 100 + 125e^{-0.12t} - 75e^{-0.04t}. \end{aligned}$$

Figure 144 shows the interesting plot of these functions. Can you give physical explanations for their main features? Why do they have the limit 100? Why is y_2 not monotone, whereas y_1 is? Why is y_1 from some time on suddenly larger than y_2? Etc. ■

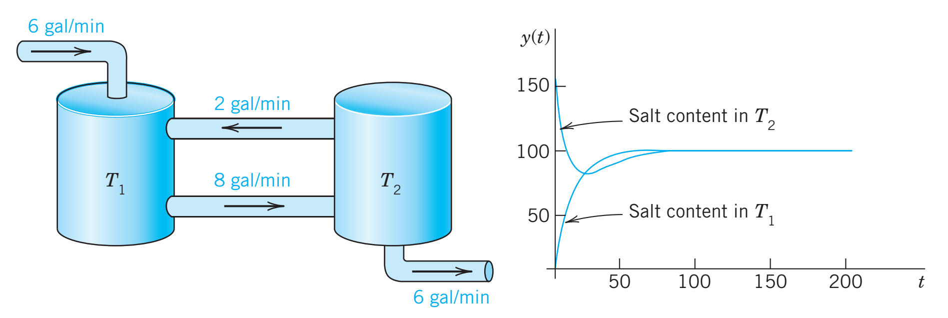


Fig. 144. Mixing problem in Example 1

Other systems of ODEs of practical importance can be solved by the Laplace transform method in a similar way, and eigenvalues and eigenvectors, as we had to determine them in Chap. 4, will come out automatically, as we have seen in Example 1.

EXAMPLE 2 Electrical Network

Find the currents $i_1(t)$ and $i_2(t)$ in the network in Fig. 145 with L and R measured in terms of the usual units (see Sec. 2.9), $v(t) = 100$ volts if $0 \leqq t \leqq 0.5$ sec and 0 thereafter, and $i(0) = 0$, $i'(0) = 0$.

Solution. The model of the network is obtained from Kirchhoff's Voltage Law as in Sec. 2.9. For the lower circuit we obtain

$$0.8i_1' + 1(i_1 - i_2) + 1.4i_1 = 100[1 - u(t - \tfrac{1}{2})]$$

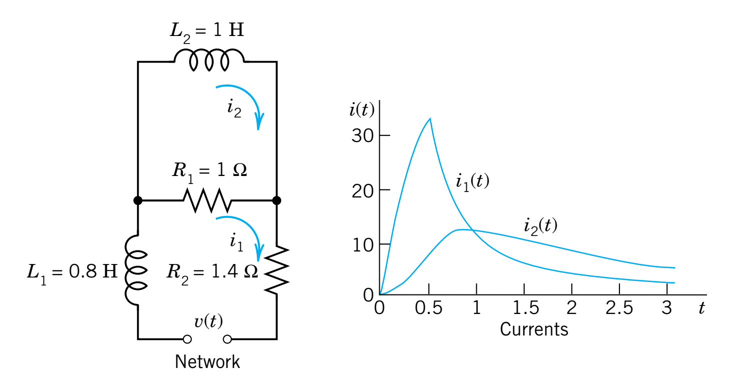


Fig. 145. Electrical network in Example 2

and for the upper

$$1 \cdot i_2' + 1(i_2 - i_1) \qquad = 0.$$

Division by 0.8 and ordering gives for the lower circuit

$$i_1' + 3i_1 - 1.25i_2 = 125[1 - u(t - \tfrac{1}{2})]$$

and for the upper

$$i_2' - i_1 + \qquad i_2 = 0.$$

With $i_1(0) = 0, i_2(0) = 0$ we obtain from (1) in Sec. 6.2 and the second shifting theorem the subsidiary system

$$\begin{aligned}(s + 3)I_1 - 1.25I_2 &= 125\left(\frac{1}{s} - \frac{e^{-s/2}}{s}\right)\\ -I_1 + (s + 1)I_2 &= 0.\end{aligned}$$

Solving algebraically for I_1 and I_2 gives

$$I_1 = \frac{125(s + 1)}{s(s + \frac{1}{2})(s + \frac{7}{2})}(1 - e^{-s/2}),$$

$$I_2 = \frac{125}{s(s + \frac{1}{2})(s + \frac{7}{2})}(1 - e^{-s/2}).$$

The right sides, without the factor $1 - e^{-s/2}$, have the partial fraction expansions

$$\frac{500}{7s} - \frac{125}{3(s + \frac{1}{2})} - \frac{625}{21(s + \frac{7}{2})}$$

and

$$\frac{500}{7s} - \frac{250}{3(s + \frac{1}{2})} + \frac{250}{21(s + \frac{7}{2})},$$

respectively. The inverse transform of this gives the solution for $0 \leqq t \leqq \frac{1}{2}$,

$$\begin{aligned} i_1(t) &= -\tfrac{125}{3}e^{-t/2} - \tfrac{625}{21}e^{-7t/2} + \tfrac{500}{7}\\ i_2(t) &= -\tfrac{250}{3}e^{-t/2} + \tfrac{250}{21}e^{-7t/2} + \tfrac{500}{7}\end{aligned} \qquad (0 \leqq t \leqq \tfrac{1}{2}).$$

According to the second shifting theorem the solution for $t > \frac{1}{2}$ is $i_1(t) - i_1(t - \frac{1}{2})$ and $i_2(t) - i_2(t - \frac{1}{2})$, that is,

$$\begin{aligned} i_1(t) &= -\tfrac{125}{3}(1 - e^{1/4})e^{-t/2} - \tfrac{625}{21}(1 - e^{7/4})e^{-7t/2} \\ i_2(t) &= -\tfrac{250}{3}(1 - e^{1/4})e^{-t/2} + \tfrac{250}{21}(1 - e^{7/4})e^{-7t/2} \end{aligned} \qquad (t > \tfrac{1}{2}).$$

Can you explain physically why both currents eventually go to zero, and why $i_1(t)$ has a sharp cusp whereas $i_2(t)$ has a continuous tangent direction at $t = \frac{1}{2}$? ■

Systems of ODEs of higher order can be solved by the Laplace transform method in a similar fashion. As an important application, typical of many similar mechanical systems, we consider coupled vibrating masses on springs.

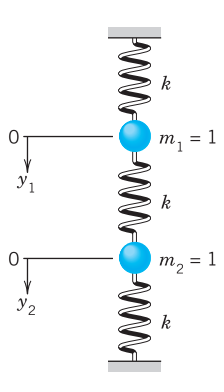


Fig. 146. Example 3

EXAMPLE 3 Model of Two Masses on Springs (Fig. 146)

The mechanical system in Fig. 146 consists of two bodies of mass 1 on three springs of the same spring constant k and of negligibly small masses of the springs. Also damping is assumed to be practically zero. Then the model of the physical system is the system of ODEs

$$\begin{aligned} y_1'' &= -ky_1 + k(y_2 - y_1) \\ y_2'' &= -k(y_2 - y_1) - ky_2. \end{aligned} \tag{3}$$

Here y_1 and y_2 are the displacements of the bodies from their positions of static equilibrium. These ODEs follow from **Newton's second law**, *Mass* × *Acceleration* = *Force,* as in Sec. 2.4 for a single body. We again regard downward forces as positive and upward as negative. On the upper body, $-ky_1$ is the force of the upper spring and $k(y_2 - y_1)$ that of the middle spring, $y_2 - y_1$ being the net change in spring length—think this over before going on. On the lower body, $-k(y_2 - y_1)$ is the force of the middle spring and $-ky_2$ that of the lower spring.

We shall determine the solution corresponding to the initial conditions $y_1(0) = 1$, $y_2(0) = 1$, $y_1'(0) = \sqrt{3k}$, $y_2'(0) = -\sqrt{3k}$. Let $Y_1 = \mathcal{L}(y_1)$ and $Y_2 = \mathcal{L}(y_2)$. Then from (2) in Sec. 6.2 and the initial conditions we obtain the subsidiary system

$$\begin{aligned} s^2Y_1 - s - \sqrt{3k} &= -kY_1 + k(Y_2 - Y_1) \\ s^2Y_2 - s + \sqrt{3k} &= -k(Y_2 - Y_1) - kY_2. \end{aligned}$$

This system of linear algebraic equations in the unknowns Y_1 and Y_2 may be written

$$\begin{aligned} (s^2 + 2k)Y_1 - \quad kY_2 \quad &= s + \sqrt{3k} \\ -ky_1 \quad + (s^2 + 2k)Y_2 &= s - \sqrt{3k}. \end{aligned}$$

Elimination (or Cramer's rule in Sec. 7.7) yields the solution, which we can expand in terms of partial fractions,

$$Y_1 = \frac{(s + \sqrt{3k})(s^2 + 2k) + k(s - \sqrt{3k})}{(s^2 + 2k)^2 - k^2} = \frac{s}{s^2 + k} + \frac{\sqrt{3k}}{s^2 + 3k}$$

$$Y_2 = \frac{(s^2 + 2k)(s - \sqrt{3k}) + k(s + \sqrt{3k})}{(s^2 + 2k)^2 - k^2} = \frac{s}{s^2 + k} - \frac{\sqrt{3k}}{s^2 + 3k}.$$

Hence the solution of our initial value problem is (Fig. 147)

$$y_1(t) = \mathcal{L}^{-1}(Y_1) = \cos\sqrt{k}t + \sin\sqrt{3k}t$$

$$y_2(t) = \mathcal{L}^{-1}(Y_2) = \cos\sqrt{k}t - \sin\sqrt{3k}t.$$

We see that the motion of each mass is harmonic (the system is undamped!), being the superposition of a "slow" oscillation and a "rapid" oscillation. ∎

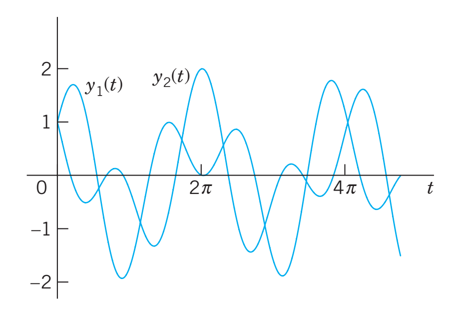


Fig. 147. Solutions in Example 3

PROBLEM SET 6.7

1. TEAM PROJECT. Comparison of Methods for Linear Systems of ODEs

(a) Models. Solve the models in Examples 1 and 2 of Sec. 4.1 by Laplace transforms and compare the amount of work with that in Sec. 4.1. Show the details of your work.

(b) Homogeneous Systems. Solve the systems (8), (11)–(13) in Sec. 4.3 by Laplace transforms. Show the details.

(c) Nonhomogeneous System. Solve the system (3) in Sec. 4.6 by Laplace transforms. Show the details.

2–15 SYSTEMS OF ODES

Using the Laplace transform and showing the details of your work, solve the IVP:

2. $y_1' + y_2 = 0,\quad y_1 + y_2' = 2\cos t,$
$y_1(0) = 1,\quad y_2(0) = 0$

3. $y_1' = -y_1 + 4y_2,\quad y_2' = 3y_1 - 2y_2,$
$y_1(0) = 3,\quad y_2(0) = 4$

4. $y_1' = 4y_2 - 8\cos 4t,\quad y_2' = -3y_1 - 9\sin 4t,$
$y_1(0) = 0,\quad y_2(0) = 3$

5. $y_1' = y_2 + 1 - u(t-1),\quad y_2' = -y_1 + 1 - u(t-1),$
$y_1(0) = 0,\quad y_2(0) = 0$

6. $y_1' = 5y_1 + y_2,\quad y_2' = y_1 + 5y_2,$
$y_1(0) = 1,\quad y_2(0) = -3$

7. $y_1' = 2y_1 - 4y_2 + u(t-1)e^t,$
$y_2' = y_1 - 3y_2 + u(t-1)e^t,\quad y_1(0) = 3,\quad y_2(0) = 0$

8. $y_1' = -2y_1 + 3y_2,\quad y_2' = 4y_1 - y_2,$
$y_1(0) = 4,\quad y_2(0) = 3$

9. $y_1' = 4y_1 + y_2,\quad y_2' = -y_1 + 2y_2,\quad y_1(0) = 3,$
$y_2(0) = 1$

10. $y_1' = -y_2,\quad y_2' = -y_1 + 2[1 - u(t - 2\pi)]\cos t,$
$y_1(0) = 1,\quad y_2(0) = 0$

11. $y_1'' = y_1 + 3y_2,\quad y_2'' = 4y_1 - 4e^t,$
$y_1(0) = 2,\quad y_1'(0) = 3,\quad y_2(0) = 1,\quad y_2'(0) = 2$

12. $y_1'' = -2y_1 + 2y_2,\quad y_2'' = 2y_1 - 5y_2,$
$y_1(0) = 1,\quad y_1'(0) = 0,\quad y_2(0) = 3,\quad y_2'(0) = 0$

13. $y_1'' + y_2 = -101\sin 10t,\quad y_2'' + y_1 = 101\sin 10t,$
$y_1(0) = 0,\quad y_1'(0) = 6,\quad y_2(0) = 8,\quad y_2'(0) = -6$

14. $4y_1' + y_2' - 2y_3' = 0, \quad -2y_1' + y_3' = 1,$
$2y_2' - 4y_3' = -16t$
$y_1(0) = 2, \quad y_2(0) = 0, \quad y_3(0) = 0$

15. $y_1' + y_2' = 2\sinh t, \quad y_2' + y_3' = e^t,$
$y_3' + y_1' = 2e^t + e^{-t}, \quad y_1(0) = 1, \quad y_2(0) = 1,$
$y_3(0) = 0$

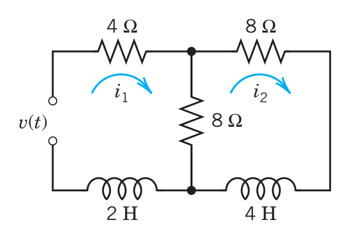
FURTHER APPLICATIONS

16. Forced vibrations of two masses. Solve the model in Example 3 with $k = 4$ and initial conditions $y_1(0) = 1$, $y_1'(0) = 1$, $y_2(0) = 1$, $y_2' = -1$ under the assumption that the force $11 \sin t$ is acting on the first body and the force $-11 \sin t$ on the second. Graph the two curves on common axes and explain the motion physically.

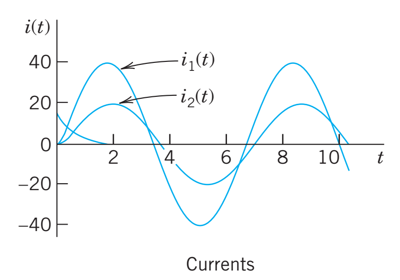
17. CAS Experiment. Effect of Initial Conditions. In Prob. 16, vary the initial conditions systematically, describe and explain the graphs physically. The great variety of curves will surprise you. Are they always periodic? Can you find empirical laws for the changes in terms of continuous changes of those conditions?

18. Mixing problem. What will happen in Example 1 if you double all flows (in particular, an increase to 12 gal/min containing 12 lb of salt from the outside), leaving the size of the tanks and the initial conditions as before? First guess, then calculate. Can you relate the new solution to the old one?

19. Electrical network. Using Laplace transforms, find the currents $i_1(t)$ and $i_2(t)$ in Fig. 148, where $v(t) = 390 \cos t$ and $i_1(0) = 0$, $i_2(0) = 0$. How soon will the currents practically reach their steady state?



Network



Currents

Fig. 148. Electrical network and currents in Problem 19

20. Single cosine wave. Solve Prob. 19 when the EMF (electromotive force) is acting from 0 to 2π only. Can you do this just by looking at Prob. 19, practically without calculation?

6.8 Laplace Transform: General Formulas

Formula	Name, Comments	Sec.
$F(s) = \mathcal{L}\{f(t)\} = \int_0^\infty e^{-st} f(t)\, dt$ $f(t) = \mathcal{L}^{-1}\{F(s)\}$	Definition of Transform Inverse Transform	6.1
$\mathcal{L}\{af(t) + bg(t)\} = a\mathcal{L}\{f(t)\} + b\mathcal{L}\{g(t)\}$	Linearity	6.1
$\mathcal{L}\{e^{at} f(t)\} = F(s - a)$ $\mathcal{L}^{-1}\{F(s - a)\} = e^{at} f(t)$	*s*-Shifting (First Shifting Theorem)	6.1
$\mathcal{L}(f') = s\mathcal{L}(f) - f(0)$ $\mathcal{L}(f'') = s^2\mathcal{L}(f) - sf(0) - f'(0)$ $\mathcal{L}(f^{(n)}) = s^n\mathcal{L}(f) - s^{(n-1)}f(0) - \cdots$ $\cdots - f^{(n-1)}(0)$ $\mathcal{L}\left\{\int_0^t f(\tau)\, d\tau\right\} = \frac{1}{s}\mathcal{L}(f)$	Differentiation of Function Integration of Function	6.2
$(f * g)(t) = \int_0^t f(\tau) g(t - \tau)\, d\tau$ $= \int_0^t f(t - \tau) g(\tau)\, d\tau$ $\mathcal{L}(f * g) = \mathcal{L}(f)\mathcal{L}(g)$	Convolution	6.5
$\mathcal{L}\{f(t - a)\, u(t - a)\} = e^{-as} F(s)$ $\mathcal{L}^{-1}\{e^{-as} F(s)\} = f(t - a)\, u(t - a)$	*t*-Shifting (Second Shifting Theorem)	6.3
$\mathcal{L}\{tf(t)\} = -F'(s)$ $\mathcal{L}\left\{\frac{f(t)}{t}\right\} = \int_s^\infty F(\tilde{s})\, d\tilde{s}$	Differentiation of Transform Integration of Transform	6.6
$\mathcal{L}(f) = \frac{1}{1 - e^{-ps}} \int_0^p e^{-st} f(t)\, dt$	*f* Periodic with Period *p*	6.4 Project 16

6.9 Table of Laplace Transforms

For more extensive tables, see Ref.* [A9] *in Appendix 1.

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
1	$1/s$	1	6.1
2	$1/s^2$	t	
3	$1/s^n \quad (n = 1, 2, \cdots)$	$t^{n-1}/(n-1)!$	
4	$1/\sqrt{s}$	$1/\sqrt{\pi t}$	
5	$1/s^{3/2}$	$2\sqrt{t/\pi}$	
6	$1/s^a \quad (a > 0)$	$t^{a-1}/\Gamma(a)$	
7	$\dfrac{1}{s-a}$	e^{at}	6.1
8	$\dfrac{1}{(s-a)^2}$	te^{at}	
9	$\dfrac{1}{(s-a)^n} \quad (n = 1, 2, \cdots)$	$\dfrac{1}{(n-1)!}t^{n-1}e^{at}$	
10	$\dfrac{1}{(s-a)^k} \quad (k > 0)$	$\dfrac{1}{\Gamma(k)}t^{k-1}e^{at}$	
11	$\dfrac{1}{(s-a)(s-b)} \quad (a \neq b)$	$\dfrac{1}{a-b}(e^{at} - e^{bt})$	
12	$\dfrac{s}{(s-a)(s-b)} \quad (a \neq b)$	$\dfrac{1}{a-b}(ae^{at} - be^{bt})$	
13	$\dfrac{1}{s^2 + \omega^2}$	$\dfrac{1}{\omega}\sin \omega t$	6.1
14	$\dfrac{s}{s^2 + \omega^2}$	$\cos \omega t$	
15	$\dfrac{1}{s^2 - a^2}$	$\dfrac{1}{a}\sinh at$	
16	$\dfrac{s}{s^2 - a^2}$	$\cosh at$	
17	$\dfrac{1}{(s-a)^2 + \omega^2}$	$\dfrac{1}{\omega}e^{at}\sinh \omega t$	
18	$\dfrac{s-a}{(s-a)^2 + \omega^2}$	$e^{at}\cos \omega t$	
19	$\dfrac{1}{s(s^2 + \omega^2)}$	$\dfrac{1}{\omega^2}(1 - \cos \omega t)$	6.2
20	$\dfrac{1}{s^2(s^2 + \omega^2)}$	$\dfrac{1}{\omega^3}(\omega t - \sin \omega t)$	

(continued)

Table of Laplace Transforms (*continued*)

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
21	$\dfrac{1}{(s^2+\omega^2)^2}$	$\dfrac{1}{2\omega^3}(\sin \omega t - \omega t \cos \omega t)$	6.6
22	$\dfrac{s}{(s^2+\omega^2)^2}$	$\dfrac{t}{2\omega} \sin \omega t$	6.6
23	$\dfrac{s^2}{(s^2+\omega^2)^2}$	$\dfrac{1}{2\omega}(\sin \omega t + \omega t \cos \omega t)$	6.6
24	$\dfrac{s}{(s^2+a^2)(s^2+b^2)} \quad (a^2 \neq b^2)$	$\dfrac{1}{b^2-a^2}(\cos at - \cos bt)$	
25	$\dfrac{1}{s^4+4k^4}$	$\dfrac{1}{4k^3}(\sin kt \cos kt - \cos kt \sinh kt)$	
26	$\dfrac{s}{s^4+4k^4}$	$\dfrac{1}{2k^2} \sin kt \sinh kt$	
27	$\dfrac{1}{s^4-k^4}$	$\dfrac{1}{2k^3}(\sinh kt - \sin kt)$	
28	$\dfrac{s}{s^4-k^4}$	$\dfrac{1}{2k^2}(\cosh kt - \cos kt)$	
29	$\sqrt{s-a} - \sqrt{s-b}$	$\dfrac{1}{2\sqrt{\pi t^3}}(e^{bt} - e^{at})$	
30	$\dfrac{1}{\sqrt{s+a}\sqrt{s+b}}$	$e^{-(a+b)t/2} I_0\left(\dfrac{a-b}{2}t\right)$	I 5.5
31	$\dfrac{1}{\sqrt{s^2+a^2}}$	$J_0(at)$	J 5.4
32	$\dfrac{s}{(s-a)^{3/2}}$	$\dfrac{1}{\sqrt{\pi t}} e^{at}(1+2at)$	
33	$\dfrac{1}{(s^2-a^2)^k} \quad (k > 0)$	$\dfrac{\sqrt{\pi}}{\Gamma(k)}\left(\dfrac{t}{2a}\right)^{k-1/2} I_{k-1/2}(at)$	I 5.5
34	e^{-as}/s	$u(t-a)$	6.3
35	e^{-as}	$\delta(t-a)$	6.4
36	$\dfrac{1}{s}e^{-k/s}$	$J_0(2\sqrt{kt})$	J 5.4
37	$\dfrac{1}{\sqrt{s}}e^{-k/s}$	$\dfrac{1}{\sqrt{\pi t}} \cos 2\sqrt{kt}$	
38	$\dfrac{1}{s^{3/2}}e^{k/s}$	$\dfrac{1}{\sqrt{\pi k}} \sinh 2\sqrt{kt}$	
39	$e^{-k\sqrt{s}} \quad (k > 0)$	$\dfrac{k}{2\sqrt{\pi t^3}} e^{-k^2/4t}$	

(*continued*)

Table of Laplace Transforms (*continued*)

	$F(s) = \mathcal{L}\{f(t)\}$	$f(t)$	Sec.
40	$\frac{1}{s} \ln s$	$-\ln t - \gamma \quad (\gamma \approx 0.5772)$	γ 5.5
41	$\ln \frac{s-a}{s-b}$	$\frac{1}{t}(e^{bt} - e^{at})$	
42	$\ln \frac{s^2 + \omega^2}{s^2}$	$\frac{2}{t}(1 - \cos \omega t)$	6.6
43	$\ln \frac{s^2 - a^2}{s^2}$	$\frac{2}{t}(1 - \cosh at)$	
44	$\arctan \frac{\omega}{s}$	$\frac{1}{t} \sin \omega t$	
45	$\frac{1}{s} \operatorname{arccot} s$	$\mathrm{Si}(t)$	App. A3.1

CHAPTER 6 REVIEW QUESTIONS AND PROBLEMS

1. State the Laplace transforms of a few simple functions from memory.

2. What are the steps of solving an ODE by the Laplace transform?

3. In what cases of solving ODEs is the present method preferable to that in Chap. 2?

4. What property of the Laplace transform is crucial in solving ODEs?

5. Is $\mathcal{L}\{f(t) + g(t)\} = \mathcal{L}\{f(t)\} + \mathcal{L}\{g(t)\}$? $\mathcal{L}\{f(t)g(t)\} = \mathcal{L}\{f(t)\}\mathcal{L}\{g(t)\}$? Explain.

6. When and how do you use the unit step function and Dirac's delta?

7. If you know $f(t) = \mathcal{L}^{-1}\{F(s)\}$, how would you find $\mathcal{L}^{-1}\{F(s)/s^2\}$?

8. Explain the use of the two shifting theorems from memory.

9. Can a discontinuous function have a Laplace transform? Give reason.

10. If two different continuous functions have transforms, the latter are different. Why is this practically important?

11–19 LAPLACE TRANSFORMS

Find the transform, indicating the method used and showing the details.

11. $5 \cosh 2t - 3 \sinh t$

12. $e^{-t}(\cos 4t - 2 \sin 4t)$

13. $\sin^2 (\frac{1}{2}\pi t)$

14. $16t^2 u(t - \frac{1}{4})$

15. $e^{t/2} u(t - 3)$

16. $u(t - 2\pi) \sin t$

17. $t \cos t + \sin t$

18. $(\sin \omega t) * (\cos \omega t)$

19. $12t * e^{-3t}$

20–28 INVERSE LAPLACE TRANSFORM

Find the inverse transform, indicating the method used and showing the details:

20. $\frac{7.5}{s^2 - 2s - 8}$

21. $\frac{s + 1}{s^2} e^{-s}$

22. $\frac{\frac{1}{16}}{s^2 + s + \frac{1}{2}}$

23. $\frac{\omega \cos \theta + s \sin \theta}{s^2 + \omega^2}$

24. $\frac{s^2 - 6.25}{(s^2 + 6.25)^2}$

25. $\frac{6(s + 1)}{s^4}$

26. $\frac{2s - 10}{s^3} e^{-5s}$

27. $\frac{3s + 4}{s^2 + 4s + 5}$

28. $\frac{3s}{s^2 - 2s + 2}$

29–37 ODEs AND SYSTEMS

Solve by the Laplace transform, showing the details and graphing the solution:

29. $y'' + 4y' + 5y = 50t, \quad y(0) = 5, \quad y'(0) = -5$

30. $y'' + 16y = 4\delta(t - \pi), \quad y(0) = -1, \quad y'(0) = 0$

31. $y'' - y' - 2y = 12u(t - \pi)\sin t, \quad y(0) = 1,$
$y'(0) = -1$

32. $y'' + 4y = \delta(t - \pi) - \delta(t - 2\pi), \quad y(0) = 1,$
$y'(0) = 0$

33. $y'' + 3y' + 2y = 2u(t - 2), \quad y(0) = 0, \quad y'(0) = 0$

34. $y_1' = y_2, \quad y_2' = -4y_1 + \delta(t - \pi), \quad y_1(0) = 0,$
$y_2(0) = 0$

35. $y_1' = 2y_1 - 4y_2, \quad y_2' = y_1 - 3y_2, \quad y_1(0) = 3,$
$y_2(0) = 0$

36. $y_1' = 2y_1 + 4y_2, \quad y_2' = y_1 + 2y_2, \quad y_1(0) = -4,$
$y_2(0) = -4$

37. $y_1' = y_2 + u(t - \pi), \quad y_2' = -y_1 + u(t - 2\pi),$
$y_1(0) = 1, \quad y_2(0) = 0$

38–45 MASS–SPRING SYSTEMS, CIRCUITS, NETWORKS

Model and solve by the Laplace transform:

38. Show that the model of the mechanical system in Fig. 149 (no friction, no damping) is

$$m_1 y_1'' = -k_1 y_1 + k_2(y_2 - y_1)$$
$$m_2 y_2'' = -k_2(y_2 - y_1) - k_3 y_2.$$

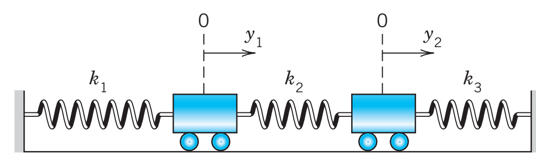


Fig. 149. System in Problems 38 and 39

39. In Prob. 38, let $m_1 = m_2 = 10$ kg, $k_1 = k_3 = 20$ kg/sec^2, $k_2 = 40$ kg/sec^2. Find the solution satisfying the initial conditions $y_1(0) = y_2(0) = 0$, $y_1'(0) = 1$ meter/sec, $y_2'(0) = -1$ meter/sec.

40. Find the model (the system of ODEs) in Prob. 38 extended by adding another mass m_3 and another spring of modulus k_4 in series.

41. Find the current $i(t)$ in the RC-circuit in Fig. 150, where $R = 10\ \Omega$, $C = 0.1$ F, $v(t) = 10t$ V if $0 < t < 4$, $v(t) = 40$ V if $t > 4$, and the initial charge on the capacitor is 0.

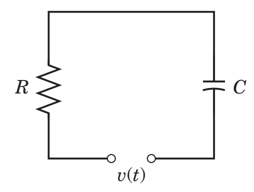


Fig. 150. RC-circuit

42. Find and graph the charge $q(t)$ and the current $i(t)$ in the LC-circuit in Fig. 151, assuming $L = 1$ H, $C = 1$ F, $v(t) = 1 - e^{-t}$ if $0 < t < \pi$, $v(t) = 0$ if $t > \pi$, and zero initial current and charge.

43. Find the current $i(t)$ in the RLC-circuit in Fig. 152, where $R = 160\ \Omega$, $L = 20$ H, $C = 0.002$ F, $v(t) = 37 \sin 10t$ V, and current and charge at $t = 0$ are zero.

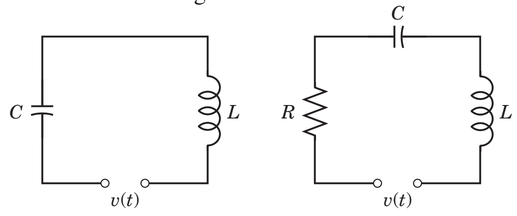


Fig. 151. LC-circuit **Fig. 152.** RLC-circuit

44. Show that, by Kirchhoff's Voltage Law (Sec. 2.9), the currents in the network in Fig. 153 are obtained from the system

$$Li_1' + R(i_1 - i_2) = v(t)$$
$$R(i_2' - i_1') + \frac{1}{C} i_2 = 0.$$

Solve this system, assuming that $R = 10\ \Omega$, $L = 20$ H, $C = 0.05$ F, $v = 20$ V, $i_1(0) = 0$, $i_2(0) = 2$ A.

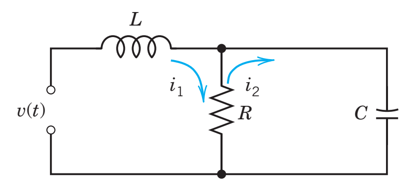


Fig. 153. Network in Problem 44

45. Set up the model of the network in Fig. 154 and find the solution, assuming that all charges and currents are 0 when the switch is closed at $t = 0$. Find the limits of $i_1(t)$ and $i_2(t)$ as $t \to \infty$, (i) from the solution, (ii) directly from the given network.

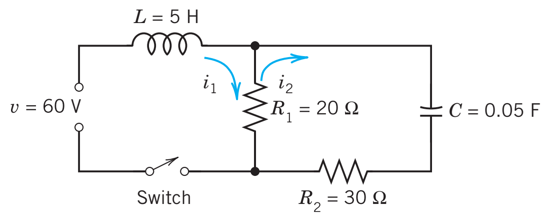


Fig. 154. Network in Problem 45

SUMMARY OF CHAPTER 6

Laplace Transforms

The main purpose of Laplace transforms is the solution of differential equations and systems of such equations, as well as corresponding initial value problems. The **Laplace transform** $F(s) = \mathcal{L}(f)$ of a function $f(t)$ is defined by

$$F(s) = \mathcal{L}(f) = \int_0^\infty e^{-st} f(t)\, dt \qquad \text{(Sec. 6.1).} \tag{1}$$

This definition is motivated by the property that the differentiation of f with respect to t corresponds to the multiplication of the transform F by s; more precisely,

$$\begin{aligned} \mathcal{L}(f') &= s\mathcal{L}(f) - f(0) \\ \mathcal{L}(f'') &= s^2\mathcal{L}(f) - sf(0) - f'(0) \end{aligned} \qquad \text{(Sec. 6.2)} \tag{2}$$

etc. Hence by taking the transform of a given differential equation

$$y'' + ay' + by = r(t) \qquad (a,\ b \text{ constant}) \tag{3}$$

and writing $\mathcal{L}(y) = Y(s)$, we obtain the **subsidiary equation**

$$(s^2 + as + b)Y = \mathcal{L}(r) + sf(0) + f'(0) + af(0). \tag{4}$$

Here, in obtaining the transform $\mathcal{L}(r)$ we can get help from the small table in Sec. 6.1 or the larger table in Sec. 6.9. This is the first step. In the second step we solve the subsidiary equation *algebraically* for $Y(s)$. In the third step we determine the **inverse transform** $y(t) = \mathcal{L}^{-1}(Y)$, that is, the solution of the problem. This is generally the hardest step, and in it we may again use one of those two tables. $Y(s)$ will often be a rational function, so that we can obtain the inverse $\mathcal{L}^{-1}(Y)$ by partial fraction reduction (Sec. 6.4) if we see no simpler way.

The Laplace method avoids the determination of a general solution of the homogeneous ODE, and we also need not determine values of arbitrary constants in a general solution from initial conditions; instead, we can insert the latter directly into (4). Two further facts account for the practical importance of the Laplace transform. First, it has some basic properties and resulting techniques that simplify the determination of transforms and inverses. The most important of these properties are listed in Sec. 6.8, together with references to the corresponding sections. More on the use of unit step functions and Dirac's delta can be found in Secs. 6.3 and 6.4, and more on convolution in Sec. 6.5. Second, due to these properties, the present method is particularly suitable for handling right sides $r(t)$ given by different expressions over different intervals of time, for instance, when $r(t)$ is a square wave or an impulse or of a form such as $r(t) = \cos t$ if $0 \leqq t \leqq 4\pi$ and 0 elsewhere.

The application of the Laplace transform to systems of ODEs is shown in Sec. 6.7. (The application to PDEs follows in Sec. 12.12.)

PART B

Linear Algebra. Vector Calculus

Matrices and *vectors*, which underlie **linear algebra** (Chaps. 7 and 8), allow us to represent numbers or functions in an ordered and compact form. Matrices can hold enormous amounts of data—think of a network of millions of computer connections or cell phone connections—in a form that can be rapidly processed by computers. The main topic of Chap. 7 is *how to solve systems of linear equations using matrices*. Concepts of rank, basis, linear transformations, and vector spaces are closely related. Chapter 8 deals with eigenvalue problems. Linear algebra is an active field that has many applications in engineering physics, numerics (see Chaps. 20–22), economics, and others.

Chapters 9 and 10 extend calculus to ***vector calculus***. We start with vectors from linear algebra and develop vector differential calculus. We differentiate functions of several variables and discuss vector differential operations such as grad, div, and curl. Chapter 10 extends regular integration to integration over curves, surfaces, and solids, thereby obtaining new types of integrals. Ingenious theorems by Gauss, Green, and Stokes allow us to transform these integrals into one another.

Software suitable for linear algebra (Lapack, Maple, Mathematica, Matlab) can be found in the list at the opening of Part E of the book if needed.

Numeric linear algebra (Chap. 20) ***can be studied directly after Chap. 7 or 8*** because Chap. 20 is independent of the other chapters in Part E on numerics.

CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants. Linear Systems

Linear algebra is a fairly extensive subject that covers vectors and matrices, determinants, systems of linear equations, vector spaces and linear transformations, eigenvalue problems, and other topics. As an area of study it has a broad appeal in that it has many applications in engineering, physics, geometry, computer science, economics, and other areas. It also contributes to a deeper understanding of mathematics itself.

Matrices, which are rectangular arrays of numbers or functions, and **vectors** are the main tools of linear algebra. Matrices are important because they let us express large amounts of data and functions in an organized and concise form. Furthermore, since matrices are single objects, we denote them by single letters and calculate with them directly. All these features have made matrices and vectors very popular for expressing scientific and mathematical ideas.

The chapter keeps a good mix between applications (electric networks, Markov processes, traffic flow, etc.) and theory. Chapter 7 is structured as follows: Sections 7.1 and 7.2 provide an intuitive introduction to matrices and vectors and their operations, including matrix multiplication. The next block of sections, that is, Secs. 7.3–7.5 provide the most important method for ***solving systems of linear equations*** by the Gauss elimination method. This method is a cornerstone of linear algebra, and the method itself and variants of it appear in different areas of mathematics and in many applications. It leads to a consideration of the behavior of solutions and concepts such as rank of a matrix, linear independence, and bases. We shift to determinants, a topic that has declined in importance, in Secs. 7.6 and 7.7. Section 7.8 covers inverses of matrices. The chapter ends with vector spaces, inner product spaces, linear transformations, and composition of linear transformations. Eigenvalue problems follow in Chap. 8.

COMMENT. ***Numeric linear algebra (Secs. 20.1–20.5) can be studied immediately after this chapter.***

Prerequisite: None.
Sections that may be omitted in a short course: 7.5, 7.9.
References and Answers to Problems: App. 1 Part B, and App. 2.

7.1 Matrices, Vectors: Addition and Scalar Multiplication

The basic concepts and rules of matrix and vector algebra are introduced in Secs. 7.1 and 7.2 and are followed by **linear systems** (systems of linear equations), a main application, in Sec. 7.3.

Let us first take a leisurely look at matrices before we formalize our discussion. A **matrix** is a rectangular array of numbers or functions which we will enclose in brackets. For example,

$$(1)\qquad \begin{bmatrix} 0.3 & 1 & -5 \\ 0 & -0.2 & 16 \end{bmatrix}, \quad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

$$\begin{bmatrix} e^{-x} & 2x^2 \\ e^{6x} & 4x \end{bmatrix}, \quad [a_1 \quad a_2 \quad a_3], \quad \begin{bmatrix} 4 \\ \frac{1}{2} \end{bmatrix}$$

are matrices. The numbers (or functions) are called **entries** or, less commonly, *elements* of the matrix. The first matrix in (1) has two **rows**, which are the horizontal lines of entries. Furthermore, it has three **columns**, which are the vertical lines of entries. The second and third matrices are **square matrices**, which means that each has as many rows as columns—3 and 2, respectively. The entries of the second matrix have two indices, signifying their location within the matrix. The first index is the number of the row and the second is the number of the column, so that together the entry's position is uniquely identified. For example, a_{23} (read *a two three*) is in Row 2 and Column 3, etc. The notation is standard and applies to all matrices, including those that are not square.

Matrices having just a single row or column are called **vectors**. Thus, the fourth matrix in (1) has just one row and is called a **row vector**. The last matrix in (1) has just one column and is called a **column vector**. Because the goal of the indexing of entries was to uniquely identify the position of an element within a matrix, one index suffices for vectors, whether they are row or column vectors. Thus, the third entry of the row vector in (1) is denoted by a_3.

Matrices are handy for storing and processing data in applications. Consider the following two common examples.

EXAMPLE 1 Linear Systems, a Major Application of Matrices

We are given a system of linear equations, briefly a **linear system**, such as

$$\begin{aligned} 4x_1 + 6x_2 + 9x_3 &= 6 \\ 6x_1 \qquad\;\; - 2x_3 &= 20 \\ 5x_1 - 8x_2 + \;\; x_3 &= 10 \end{aligned}$$

where x_1, x_2, x_3 are the **unknowns**. We form the **coefficient matrix**, call it **A**, by listing the coefficients of the unknowns in the position in which they appear in the linear equations. In the second equation, there is no unknown x_2, which means that the coefficient of x_2 is 0 and hence in matrix **A**, $a_{22} = 0$, Thus,

$$\mathbf{A} = \begin{bmatrix} 4 & 6 & 9 \\ 6 & 0 & -2 \\ 5 & -8 & 1 \end{bmatrix}. \quad \text{We form another matrix} \quad \tilde{\mathbf{A}} = \begin{bmatrix} 4 & 6 & 9 & 6 \\ 6 & 0 & -2 & 20 \\ 5 & -8 & 1 & 10 \end{bmatrix}$$

by augmenting **A** with the right sides of the linear system and call it the augmented matrix of the system.

Since we can go back and recapture the system of linear equations directly from the augmented matrix $\tilde{\mathbf{A}}$, $\tilde{\mathbf{A}}$ contains all the information of the system and can thus be used to solve the linear system. This means that we can just use the augmented matrix to do the calculations needed to solve the system. We shall explain this in detail in Sec. 7.3. Meanwhile you may verify by substitution that the solution is $x_1 = 3, x_2 = \frac{1}{2}, x_3 = -1$.

The notation x_1, x_2, x_3 for the unknowns is practical but not essential; we could choose x, y, z or some other letters. ∎

EXAMPLE 2 Sales Figures in Matrix Form

Sales figures for three products I, II, III in a store on Monday (Mon), Tuesday (Tues), $\cdots$ may for each week be arranged in a matrix

$$\mathbf{A} = \begin{array}{c} \text{Mon} \quad \text{Tues} \quad \text{Wed} \quad \text{Thur} \quad \text{Fri} \quad \text{Sat} \quad \text{Sun} \\ \begin{bmatrix} 40 & 33 & 81 & 0 & 21 & 47 & 33 \\ 0 & 12 & 78 & 50 & 50 & 96 & 90 \\ 10 & 0 & 0 & 27 & 43 & 78 & 56 \end{bmatrix} \end{array} \cdot \begin{array}{c} \text{I} \\ \text{II} \\ \text{III} \end{array}$$

If the company has 10 stores, we can set up 10 such matrices, one for each store. Then, by adding corresponding entries of these matrices, we can get a matrix showing the total sales of each product on each day. Can you think of other data which can be stored in matrix form? For instance, in transportation or storage problems? Or in listing distances in a network of roads? ∎

General Concepts and Notations

Let us formalize what we just have discussed. We shall denote matrices by capital boldface letters **A**, **B**, **C**, $\cdots$, or by writing the general entry in brackets; thus $\mathbf{A} = [a_{jk}]$, and so on. By an $\boldsymbol{m \times n}$ **matrix** (read *m by n matrix*) we mean a matrix with m rows and n columns—rows always come first! $m \times n$ is called the **size** of the matrix. Thus an $m \times n$ matrix is of the form

(2)
$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}.$$

The matrices in (1) are of sizes 2×3, 3×3, 2×2, 1×3, and 2×1, respectively.

Each entry in (2) has two subscripts. The first is the *row number* and the second is the *column number*. Thus a_{21} is the entry in Row 2 and Column 1.

If $m = n$, we call **A** an $n \times n$ **square matrix**. Then its diagonal containing the entries $a_{11}, a_{22}, \cdots, a_{nn}$ is called the **main diagonal** of **A**. Thus the main diagonals of the two square matrices in (1) are a_{11}, a_{22}, a_{33} and e^{-x}, $4x$, respectively.

Square matrices are particularly important, as we shall see. A matrix of any size $m \times n$ is called a **rectangular matrix**; this includes square matrices as a special case.

Vectors

A **vector** is a matrix with only one row or column. Its entries are called the **components** of the vector. We shall denote vectors by ***lowercase*** boldface letters $\mathbf{a}, \mathbf{b}, \cdots$ or by its general component in brackets, $\mathbf{a} = [a_j]$, and so on. Our special vectors in (1) suggest that a (general) **row vector** is of the form

$$\mathbf{a} = [a_1 \quad a_2 \quad \cdots \quad a_n]. \qquad \text{For instance,} \qquad \mathbf{a} = [-2 \quad 5 \quad 0.8 \quad 0 \quad 1].$$

A **column vector** is of the form

$$\mathbf{b} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}. \qquad \text{For instance,} \qquad \mathbf{b} = \begin{bmatrix} 4 \\ 0 \\ -7 \end{bmatrix}.$$

Addition and Scalar Multiplication of Matrices and Vectors

What makes matrices and vectors really useful and particularly suitable for computers is the fact that we can calculate with them almost as easily as with numbers. Indeed, we now introduce rules for addition and for scalar multiplication (multiplication by numbers) that were suggested by practical applications. (Multiplication of matrices by matrices follows in the next section.) We first need the concept of equality.

DEFINITION

Equality of Matrices

Two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ are **equal**, written $\mathbf{A} = \mathbf{B}$, if and only if they have the same size and the corresponding entries are equal, that is, $a_{11} = b_{11}$, $a_{12} = b_{12}$, and so on. Matrices that are not equal are called **different**. Thus, matrices of different sizes are always different.

EXAMPLE 3 **Equality of Matrices**

Let

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \qquad \text{and} \qquad \mathbf{B} = \begin{bmatrix} 4 & 0 \\ 3 & -1 \end{bmatrix}.$$

Then

$$\mathbf{A} = \mathbf{B} \qquad \text{if and only if} \qquad \begin{matrix} a_{11} = 4, & a_{12} = 0, \\ a_{21} = 3, & a_{22} = -1. \end{matrix}$$

The following matrices are all different. Explain!

$$\begin{bmatrix} 1 & 3 \\ 4 & 2 \end{bmatrix} \quad \begin{bmatrix} 4 & 2 \\ 1 & 3 \end{bmatrix} \quad \begin{bmatrix} 4 & 1 \\ 2 & 3 \end{bmatrix} \quad \begin{bmatrix} 1 & 3 & 0 \\ 4 & 2 & 0 \end{bmatrix} \quad \begin{bmatrix} 0 & 1 & 3 \\ 0 & 4 & 2 \end{bmatrix}$$

■

DEFINITION

Addition of Matrices

The **sum** of two matrices $\mathbf{A} = [a_{jk}]$ and $\mathbf{B} = [b_{jk}]$ ***of the same size*** is written $\mathbf{A} + \mathbf{B}$ and has the entries $a_{jk} + b_{jk}$ obtained by adding the corresponding entries of $\mathbf{A}$ and $\mathbf{B}$. Matrices of different sizes cannot be added.

As a special case, the **sum a + b** of two row vectors or two column vectors, which must have the same number of components, is obtained by adding the corresponding components.

EXAMPLE 4 **Addition of Matrices and Vectors**

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} -4 & 6 & 3 \\ 0 & 1 & 2 \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 5 & -1 & 0 \\ 3 & 1 & 0 \end{bmatrix}, \quad \text{then} \quad \mathbf{A} + \mathbf{B} = \begin{bmatrix} 1 & 5 & 3 \\ 3 & 2 & 2 \end{bmatrix}.$$

$\mathbf{A}$ in Example 3 and our present $\mathbf{A}$ cannot be added. If $\mathbf{a} = [5 \quad 7 \quad 2]$ and $\mathbf{b} = [-6 \quad 2 \quad 0]$, then $\mathbf{a} + \mathbf{b} = [-1 \quad 9 \quad 2]$.

An application of matrix addition was suggested in Example 2. Many others will follow. ■

DEFINITION

Scalar Multiplication (Multiplication by a Number)

The **product** of any $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ and any **scalar** c (number c) is written $c\mathbf{A}$ and is the $m \times n$ matrix $c\mathbf{A} = [ca_{jk}]$ obtained by multiplying each entry of $\mathbf{A}$ by c.

Here $(-1)\mathbf{A}$ is simply written $-\mathbf{A}$ and is called the **negative** of $\mathbf{A}$. Similarly, $(-k)\mathbf{A}$ is written $-k\mathbf{A}$. Also, $\mathbf{A} + (-\mathbf{B})$ is written $\mathbf{A} - \mathbf{B}$ and is called the **difference** of $\mathbf{A}$ and $\mathbf{B}$ (which must have the same size!).

EXAMPLE 5 **Scalar Multiplication**

$$\text{If} \quad \mathbf{A} = \begin{bmatrix} 2.7 & -1.8 \\ 0 & 0.9 \\ 9.0 & -4.5 \end{bmatrix}, \quad \text{then} \quad -\mathbf{A} = \begin{bmatrix} -2.7 & 1.8 \\ 0 & -0.9 \\ -9.0 & 4.5 \end{bmatrix}, \quad \frac{10}{9}\mathbf{A} = \begin{bmatrix} 3 & -2 \\ 0 & 1 \\ 10 & -5 \end{bmatrix}, \quad 0\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

If a matrix $\mathbf{B}$ shows the distances between some cities in miles, $1.609\mathbf{B}$ gives these distances in kilometers. ■

Rules for Matrix Addition and Scalar Multiplication. From the familiar laws for the addition of numbers we obtain similar laws for the addition of matrices of the same size $m \times n$, namely,

$$\textbf{(3)} \quad \begin{aligned} &\text{(a)} \quad && \mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \\ &\text{(b)} \quad && (\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \qquad (\text{written } \mathbf{A} + \mathbf{B} + \mathbf{C}) \\ &\text{(c)} \quad && \mathbf{A} + \mathbf{0} = \mathbf{A} \\ &\text{(d)} \quad && \mathbf{A} + (-\mathbf{A}) = \mathbf{0}. \end{aligned}$$

Here $\mathbf{0}$ denotes the **zero matrix** (of size $m \times n$), that is, the $m \times n$ matrix with all entries zero. If $m = 1$ or $n = 1$, this is a vector, called a **zero vector**.

Hence matrix addition is *commutative* and *associative* [by (3a) and (3b)].

Similarly, for scalar multiplication we obtain the rules

$$\textbf{(4)}\qquad \begin{aligned} &\text{(a)} \quad c(\mathbf{A}+\mathbf{B}) = c\mathbf{A} + c\mathbf{B} \\ &\text{(b)} \quad (c+k)\mathbf{A} = c\mathbf{A} + k\mathbf{A} \\ &\text{(c)} \quad c(k\mathbf{A}) = (ck)\mathbf{A} \qquad \text{(written } ck\mathbf{A}\text{)} \\ &\text{(d)} \quad 1\mathbf{A} = \mathbf{A}. \end{aligned}$$

PROBLEM SET 7.1

1–7 GENERAL QUESTIONS

1. **Equality.** Give reasons why the five matrices in Example 3 are all different.
2. **Double subscript notation.** If you write the matrix in Example 2 in the form $\mathbf{A} = [a_{jk}]$, what is a_{31}? a_{13}? a_{26}? a_{33}?
3. **Sizes.** What sizes do the matrices in Examples 1, 2, 3, and 5 have?
4. **Main diagonal.** What is the main diagonal of **A** in Example 1? Of **A** and **B** in Example 3?
5. **Scalar multiplication.** If **A** in Example 2 shows the number of items sold, what is the matrix **B** of units sold if a unit consists of **(a)** 5 items and **(b)** 10 items?
6. If a 12×12 matrix **A** shows the distances between 12 cities in kilometers, how can you obtain from **A** the matrix **B** showing these distances in miles?
7. **Addition of vectors.** Can you add: A row and a column vector with different numbers of components? With the same number of components? Two row vectors with the same number of components but different numbers of zeros? A vector and a scalar? A vector with four components and a 2×2 matrix?

8–16 ADDITION AND SCALAR MULTIPLICATION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 0 & 2 & 4 \\ 6 & 5 & 5 \\ 1 & 0 & -3 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 0 & 5 & 2 \\ 5 & 3 & 4 \\ -2 & 4 & -2 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 5 & 2 \\ -2 & 4 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{D} = \begin{bmatrix} -4 & 1 \\ 5 & 0 \\ 2 & -1 \end{bmatrix},$$

$$\mathbf{E} = \begin{bmatrix} 0 & 2 \\ 3 & 4 \\ 3 & -1 \end{bmatrix}$$

$$\mathbf{u} = \begin{bmatrix} 1.5 \\ 0 \\ -3.0 \end{bmatrix}, \qquad \mathbf{v} = \begin{bmatrix} -1 \\ 3 \\ 2 \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} -5 \\ -30 \\ 10 \end{bmatrix}.$$

Find the following expressions, indicating which of the rules in (3) or (4) they illustrate, or give reasons why they are not defined.

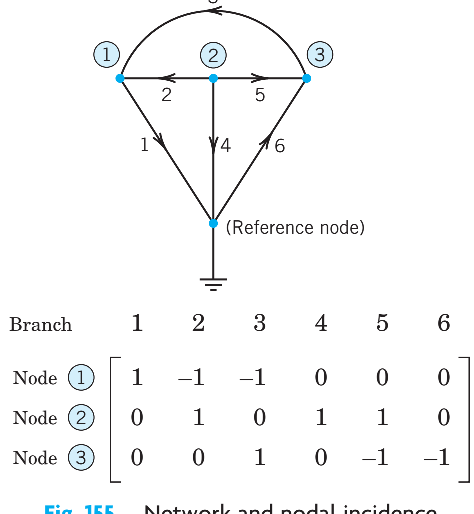
8. $2\mathbf{A} + 4\mathbf{B}$, $4\mathbf{B} + 2\mathbf{A}$, $0\mathbf{A} + \mathbf{B}$, $0.4\mathbf{B} - 4.2\mathbf{A}$
9. $3\mathbf{A}$, $0.5\mathbf{B}$, $3\mathbf{A} + 0.5\mathbf{B}$, $3\mathbf{A} + 0.5\mathbf{B} + \mathbf{C}$
10. $(4 \cdot 3)\mathbf{A}$, $4(3\mathbf{A})$, $14\mathbf{B} - 3\mathbf{B}$, $11\mathbf{B}$
11. $8\mathbf{C} + 10\mathbf{D}$, $2(5\mathbf{D} + 4\mathbf{C})$, $0.6\mathbf{C} - 0.6\mathbf{D}$, $0.6(\mathbf{C} - \mathbf{D})$
12. $(\mathbf{C} + \mathbf{D}) + \mathbf{E}$, $(\mathbf{D} + \mathbf{E}) + \mathbf{C}$, $0(\mathbf{C} - \mathbf{E}) + 4\mathbf{D}$, $\mathbf{A} - 0\mathbf{C}$
13. $(2 \cdot 7)\mathbf{C}$, $2(7\mathbf{C})$, $-\mathbf{D} + 0\mathbf{E}$, $\mathbf{E} - \mathbf{D} + \mathbf{C} + \mathbf{u}$
14. $(5\mathbf{u} + 5\mathbf{v}) - \frac{1}{2}\mathbf{w}$, $-20(\mathbf{u} + \mathbf{v}) + 2\mathbf{w}$, $\mathbf{E} - (\mathbf{u} + \mathbf{v})$, $10(\mathbf{u} + \mathbf{v}) + \mathbf{w}$
15. $(\mathbf{u} + \mathbf{v}) - \mathbf{w}$, $\mathbf{u} + (\mathbf{v} - \mathbf{w})$, $\mathbf{C} + 0\mathbf{w}$, $0\mathbf{E} + \mathbf{u} - \mathbf{v}$
16. $15\mathbf{v} - 3\mathbf{w} - 0\mathbf{u}$, $-3\mathbf{w} + 15\mathbf{v}$, $\mathbf{D} - \mathbf{u} + 3\mathbf{C}$, $8.5\mathbf{w} - 11.1\mathbf{u} + 0.4\mathbf{v}$
17. **Resultant of forces.** If the above vectors **u**, **v**, **w** represent forces in space, their sum is called their *resultant*. Calculate it.
18. **Equilibrium.** By definition, forces are *in equilibrium* if their resultant is the zero vector. Find a force **p** such that the above **u**, **v**, **w**, and **p** are in equilibrium.
19. **General rules.** Prove (3) and (4) for general 2×3 matrices and scalars c and k.

20. TEAM PROJECT. Matrices for Networks. Matrices have various engineering applications, as we shall see. For instance, they can be used to characterize connections in electrical networks, in nets of roads, in production processes, etc., as follows.

(a) Nodal Incidence Matrix. The network in Fig. 155 consists of six *branches* (connections) and four *nodes* (points where two or more branches come together). One node is the *reference node* (grounded node, whose voltage is zero). We number the other nodes and number and direct the branches. This we do arbitrarily. The network can now be described by a matrix $\mathbf{A} = [a_{jk}]$, where

$$a_{jk} = \begin{cases} +1 \text{ if branch } k \text{ leaves node } \textcircled{j} \\ -1 \text{ if branch } k \text{ enters node } \textcircled{j} \\ \;\;\,0 \text{ if branch } k \text{ does not touch node } \textcircled{j}. \end{cases}$$

$\mathbf{A}$ is called the *nodal incidence matrix* of the network. Show that for the network in Fig. 155 the matrix $\mathbf{A}$ has the given form.



Branch	1	2	3	4	5	6
Node ①	1	–1	–1	0	0	0
Node ②	0	1	0	1	1	0
Node ③	0	0	1	0	–1	–1

Fig. 155. Network and nodal incidence matrix in Team Project 20(a)

(b) Find the nodal incidence matrices of the networks in Fig. 156.

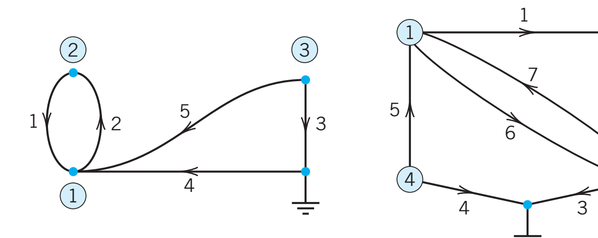


Fig. 156. Electrical networks in Team Project 20(b)

(c) Sketch the three networks corresponding to the nodal incidence matrices

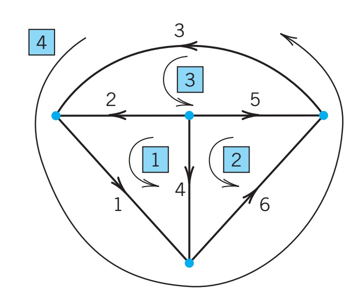
$$\begin{bmatrix} 1 & 0 & 0 & 1 \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & -1 & 0 & 0 & 1 \\ -1 & 1 & -1 & 1 & 0 \\ 0 & 0 & 1 & -1 & 0 \end{bmatrix},$$

$$\begin{bmatrix} 1 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 1 & 0 \\ 0 & -1 & -1 & 0 & 1 \end{bmatrix}.$$

(d) Mesh Incidence Matrix. A network can also be characterized by the *mesh incidence matrix* $\mathbf{M} = [m_{jk}]$, where

$$m_{jk} = \begin{cases} +1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \quad\text{and has the same orientation} \\ -1 \text{ if branch } k \text{ is in mesh } \boxed{j} \\ \quad\text{and has the opposite orientation} \\ \;\;\,0 \text{ if branch } k \text{ is not in mesh } \boxed{j} \end{cases}$$

and a mesh is a loop with no branch in its interior (or in its exterior). Here, the meshes are numbered and directed (oriented) in an arbitrary fashion. Show that for the network in Fig. 157, the matrix $\mathbf{M}$ has the given form, where Row 1 corresponds to mesh 1, etc.



$$\mathbf{M} = \begin{bmatrix} 1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 1 \\ 0 & -1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 1 \end{bmatrix}$$

Fig. 157. Network and matrix $\mathbf{M}$ in Team Project 20(d)

7.2 Matrix Multiplication

Matrix multiplication means that one multiplies matrices by matrices. Its definition is standard but it looks artificial. *Thus you have to study matrix multiplication carefully*, multiply a few matrices together for practice until you can understand how to do it. Here then is the definition. (Motivation follows later.)

DEFINITION

Multiplication of a Matrix by a Matrix

The **product** $\mathbf{C} = \mathbf{AB}$ (in this order) of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ times an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined if and only if $r = n$ and is then the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

(1) $$c_{jk} = \sum_{l=1}^{n} a_{jl}b_{lk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \begin{array}{l} j = 1, \cdots, m \\ k = 1, \cdots, p. \end{array}$$

The condition $r = n$ means that the second factor, **B**, must have as many rows as the first factor has columns, namely n. A diagram of sizes that shows when matrix multiplication is possible is as follows:

$$\begin{array}{cccc} \mathbf{A} & \mathbf{B} & = & \mathbf{C} \\ [m \times n] & [n \times p] & = & [m \times p]. \end{array}$$

The entry c_{jk} in (1) is obtained by multiplying each entry in the jth row of **A** by the corresponding entry in the kth column of **B** and then adding these n products. For instance, $c_{21} = a_{21}b_{11} + a_{22}b_{21} + \cdots + a_{2n}b_{n1}$, and so on. One calls this briefly a ***multiplication of rows into columns***. For $n = 3$, this is illustrated by

$$m = 4 \left\{ \overbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ a_{41} & a_{42} & a_{43} \end{bmatrix}}^{n=3} \overbrace{\begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}}^{p=2} = \overbrace{\begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \\ c_{31} & c_{32} \\ c_{41} & c_{42} \end{bmatrix}}^{p=2} \right\} m = 4$$

Notations in a product $\mathbf{AB} = \mathbf{C}$

where we shaded the entries that contribute to the calculation of entry c_{21} just discussed.

Matrix multiplication will be motivated by its use in *linear transformations* in this section and more fully in Sec. 7.9.

Let us illustrate the main points of matrix multiplication by some examples. Note that matrix multiplication also includes multiplying a matrix by a vector, since, after all, a vector is a special matrix.

EXAMPLE 1 **Matrix Multiplication**

$$\mathbf{AB} = \begin{bmatrix} 3 & 5 & -1 \\ 4 & 0 & 2 \\ -6 & -3 & 2 \end{bmatrix} \begin{bmatrix} 2 & -2 & 3 & 1 \\ 5 & 0 & 7 & 8 \\ 9 & -4 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 22 & -2 & 43 & 42 \\ 26 & -16 & 14 & 6 \\ -9 & 4 & -37 & -28 \end{bmatrix}$$

Here $c_{11} = 3 \cdot 2 + 5 \cdot 5 + (-1) \cdot 9 = 22$, and so on. The entry in the box is $c_{23} = 4 \cdot 3 + 0 \cdot 7 + 2 \cdot 1 = 14$. The product **BA** is not defined. ∎

EXAMPLE 2 **Multiplication of a Matrix and a Vector**

$$\begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix}\begin{bmatrix} 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 4 \cdot 3 + 2 \cdot 5 \\ 1 \cdot 3 + 8 \cdot 5 \end{bmatrix} = \begin{bmatrix} 22 \\ 43 \end{bmatrix} \qquad \text{whereas} \qquad \begin{bmatrix} 3 \\ 5 \end{bmatrix}\begin{bmatrix} 4 & 2 \\ 1 & 8 \end{bmatrix} \qquad \text{is undefined.}$$

EXAMPLE 3 **Products of Row and Column Vectors**

$$\begin{bmatrix} 3 & 6 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix} = \begin{bmatrix} 19 \end{bmatrix}, \qquad \begin{bmatrix} 1 \\ 2 \\ 4 \end{bmatrix}\begin{bmatrix} 3 & 6 & 1 \end{bmatrix} = \begin{bmatrix} 3 & 6 & 1 \\ 6 & 12 & 2 \\ 12 & 24 & 4 \end{bmatrix}.$$

EXAMPLE 4 **CAUTION! Matrix Multiplication Is Not Commutative, AB ≠ BA in General**

This is illustrated by Examples 1 and 2, where one of the two products is not even defined, and by Example 3, where the two products have different sizes. But it also holds for square matrices. For instance,

$$\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} \qquad \text{but} \qquad \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 100 & 100 \end{bmatrix} = \begin{bmatrix} 99 & 99 \\ -99 & -99 \end{bmatrix}.$$

It is interesting that this also shows that $\mathbf{AB} = \mathbf{0}$ does ***not*** necessarily imply $\mathbf{BA} = \mathbf{0}$ or $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$. We shall discuss this further in Sec. 7.8, along with reasons when this happens.

Our examples show that in matrix products ***the order of factors must always be observed very carefully***. Otherwise matrix multiplication satisfies rules similar to those for numbers, namely.

$$\text{(2)} \qquad \begin{aligned} &\text{(a)} \quad (k\mathbf{A})\mathbf{B} = k(\mathbf{AB}) = \mathbf{A}(k\mathbf{B}) \quad \textit{written } k\mathbf{AB} \textit{ or } \mathbf{A}k\mathbf{B} \\ &\text{(b)} \quad \mathbf{A}(\mathbf{BC}) = (\mathbf{AB})\mathbf{C} \qquad\qquad \textit{written } \mathbf{ABC} \\ &\text{(c)} \quad (\mathbf{A} + \mathbf{B})\mathbf{C} = \mathbf{AC} + \mathbf{BC} \\ &\text{(d)} \quad \mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB} \end{aligned}$$

provided **A**, **B**, and **C** are such that the expressions on the left are defined; here, k is any scalar. (2b) is called the **associative law**. (2c) and (2d) are called the **distributive laws**.

Since matrix multiplication is a multiplication of rows into columns, we can write the defining formula (1) more compactly as

$$\textbf{(3)} \qquad c_{jk} = \mathbf{a}_j\mathbf{b}_k, \qquad j = 1, \cdots, m; \quad k = 1, \cdots, p,$$

where $\mathbf{a}_j$ is the jth row vector of **A** and $\mathbf{b}_k$ is the kth column vector of **B**, so that in agreement with (1),

$$\mathbf{a}_j\mathbf{b}_k = \begin{bmatrix} a_{j1} & a_{j2} & \cdots & a_{jn} \end{bmatrix}\begin{bmatrix} b_{1k} \\ \vdots \\ b_{nk} \end{bmatrix} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk}.$$

EXAMPLE 5 Product in Terms of Row and Column Vectors

If $\mathbf{A} = [a_{jk}]$ is of size 3×3 and $\mathbf{B} = [b_{jk}]$ is of size 3×4, then

$$(4) \qquad \mathbf{AB} = \begin{bmatrix} \mathbf{a}_1\mathbf{b}_1 & \mathbf{a}_1\mathbf{b}_2 & \mathbf{a}_1\mathbf{b}_3 & \mathbf{a}_1\mathbf{b}_4 \\ \mathbf{a}_2\mathbf{b}_1 & \mathbf{a}_2\mathbf{b}_2 & \mathbf{a}_2\mathbf{b}_3 & \mathbf{a}_2\mathbf{b}_4 \\ \mathbf{a}_3\mathbf{b}_1 & \mathbf{a}_3\mathbf{b}_2 & \mathbf{a}_3\mathbf{b}_3 & \mathbf{a}_3\mathbf{b}_4 \end{bmatrix}.$$

Taking $\mathbf{a}_1 = [3 \quad 5 \quad -1]$, $\mathbf{a}_2 = [4 \quad 0 \quad 2]$, etc., verify (4) for the product in Example 1. ■

Parallel processing of products on the computer is facilitated by a variant of (3) for computing $\mathbf{C} = \mathbf{AB}$, which is used by standard algorithms (such as in Lapack). In this method, $\mathbf{A}$ is used as given, $\mathbf{B}$ is taken in terms of its column vectors, and the product is computed columnwise; thus,

$$(5) \qquad \mathbf{AB} = \mathbf{A}[\mathbf{b}_1 \quad \mathbf{b}_2 \quad \cdots \quad \mathbf{b}_p] = [\mathbf{Ab}_1 \quad \mathbf{Ab}_2 \quad \cdots \quad \mathbf{Ab}_p].$$

Columns of $\mathbf{B}$ are then assigned to different processors (individually or several to each processor), which simultaneously compute the columns of the product matrix $\mathbf{Ab}_1$, $\mathbf{Ab}_2$, etc.

EXAMPLE 6 Computing Products Columnwise by (5)

To obtain

$$\mathbf{AB} = \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 & 0 & 7 \\ -1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 11 & 4 & 34 \\ -17 & 8 & -23 \end{bmatrix}$$

from (5), calculate the columns

$$\begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 3 \\ -1 \end{bmatrix} = \begin{bmatrix} 11 \\ -17 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 4 \\ 8 \end{bmatrix}, \quad \begin{bmatrix} 4 & 1 \\ -5 & 2 \end{bmatrix} \begin{bmatrix} 7 \\ 6 \end{bmatrix} = \begin{bmatrix} 34 \\ -23 \end{bmatrix}$$

of $\mathbf{AB}$ and then write them as a single matrix, as shown in the first formula on the right. ■

Motivation of Multiplication by Linear Transformations

Let us now motivate the "unnatural" matrix multiplication by its use in **linear transformations**. For $n = 2$ variables these transformations are of the form

$$(6^*) \qquad \begin{aligned} y_1 &= a_{11}x_1 + a_{12}x_2 \\ y_2 &= a_{21}x_1 + a_{22}x_2 \end{aligned}$$

and suffice to explain the idea. (For general n they will be discussed in Sec. 7.9.) For instance, (6*) may relate an x_1x_2-coordinate system to a y_1y_2-coordinate system in the plane. In vectorial form we can write (6*) as

$$(6) \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{Ax} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} a_{11}x_1 + a_{12}x_2 \\ a_{21}x_1 + a_{22}x_2 \end{bmatrix}.$$

Now suppose further that the x_1x_2-system is related to a w_1w_2-system by another linear transformation, say,

$$\text{(7)}\qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \mathbf{Bw} = \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} b_{11}w_1 + b_{12}w_2 \\ b_{21}w_1 + b_{22}w_2 \end{bmatrix}.$$

Then the y_1y_2-system is related to the w_1w_2-system indirectly via the x_1x_2-system, and we wish to express this relation directly. Substitution will show that this direct relation is a linear transformation, too, say,

$$\text{(8)}\qquad \mathbf{y} = \mathbf{Cw} = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} c_{11}w_1 + c_{12}w_2 \\ c_{21}w_1 + c_{22}w_2 \end{bmatrix}.$$

Indeed, substituting (7) into (6), we obtain

$$\begin{aligned} y_1 &= a_{11}(b_{11}w_1 + b_{12}w_2) + a_{12}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{11}b_{11} + a_{12}b_{21})w_1 + (a_{11}b_{12} + a_{12}b_{22})w_2 \\ y_2 &= a_{21}(b_{11}w_1 + b_{12}w_2) + a_{22}(b_{21}w_1 + b_{22}w_2) \\ &= (a_{21}b_{11} + a_{22}b_{21})w_1 + (a_{21}b_{12} + a_{22}b_{22})w_2. \end{aligned}$$

Comparing this with (8), we see that

$$\begin{aligned} c_{11} &= a_{11}b_{11} + a_{12}b_{21} \qquad & c_{12} &= a_{11}b_{12} + a_{12}b_{22} \\ c_{21} &= a_{21}b_{11} + a_{22}b_{21} \qquad & c_{22} &= a_{21}b_{12} + a_{22}b_{22}. \end{aligned}$$

This proves that $\mathbf{C} = \mathbf{AB}$ with the product defined as in (1). For larger matrix sizes the idea and result are exactly the same. Only the number of variables changes. We then have m variables y and n variables x and p variables w. The matrices $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C} = \mathbf{AB}$ then have sizes $m \times n$, $n \times p$, and $m \times p$, respectively. And the requirement that $\mathbf{C}$ be the product $\mathbf{AB}$ leads to formula (1) in its general form. *This motivates matrix multiplication.*

Transposition

We obtain the transpose of a matrix by writing its rows as columns (or equivalently its columns as rows). This also applies to the transpose of vectors. Thus, a row vector becomes a column vector and vice versa. In addition, for square matrices, we can also "reflect" the elements along the main diagonal, that is, interchange entries that are symmetrically positioned with respect to the main diagonal to obtain the transpose. Hence a_{12} becomes a_{21}, a_{31} becomes a_{13}, and so forth. Example 7 illustrates these ideas. Also note that, if $\mathbf{A}$ is the given matrix, then we denote its transpose by $\mathbf{A}^\mathsf{T}$.

EXAMPLE 7 Transposition of Matrices and Vectors

If $\qquad \mathbf{A} = \begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}, \qquad$ then $\qquad \mathbf{A}^\mathsf{T} = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}.$

A little more compactly, we can write

$$\begin{bmatrix} 5 & -8 & 1 \\ 4 & 0 & 0 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 5 & 4 \\ -8 & 0 \\ 1 & 0 \end{bmatrix}, \qquad \begin{bmatrix} 3 & 0 & 7 \\ 8 & -1 & 5 \\ 1 & -9 & 4 \end{bmatrix}^{\mathsf{T}} = \begin{bmatrix} 3 & 8 & 1 \\ 0 & -1 & -9 \\ 7 & 5 & 4 \end{bmatrix},$$

Furthermore, the transpose $[6 \quad 2 \quad 3]^{\mathsf{T}}$ of the row vector $[6 \quad 2 \quad 3]$ is the column vector

$$[6 \quad 2 \quad 3]^{\mathsf{T}} = \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix} \cdot \qquad \text{Conversely,} \qquad \begin{bmatrix} 6 \\ 2 \\ 3 \end{bmatrix}^{\mathsf{T}} = [6 \quad 2 \quad 3].$$

DEFINITION

Transposition of Matrices and Vectors

The transpose of an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the $n \times m$ matrix $\mathbf{A}^{\mathsf{T}}$ (read *A transpose*) that has the first *row* of $\mathbf{A}$ as its first *column*, the second *row* of $\mathbf{A}$ as its second *column*, and so on. Thus the transpose of $\mathbf{A}$ in (2) is $\mathbf{A}^{\mathsf{T}} = [a_{kj}]$, written out

$$\mathbf{A}^{\mathsf{T}} = [a_{kj}] = \begin{bmatrix} a_{11} & a_{21} & \cdots & a_{m1} \\ a_{12} & a_{22} & \cdots & a_{m2} \\ \cdot & \cdot & \cdots & \cdot \\ a_{1n} & a_{2n} & \cdots & a_{mn} \end{bmatrix}. \tag{9}$$

As a special case, transposition converts row vectors to column vectors and conversely.

Transposition gives us a choice in that we can work either with the matrix or its transpose, whichever is more convenient.

Rules for transposition are

$$\begin{aligned} &\text{(a)} \quad (\mathbf{A}^{\mathsf{T}})^{\mathsf{T}} = \mathbf{A} \\ &\text{(b)} \quad (\mathbf{A} + \mathbf{B})^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} + \mathbf{B}^{\mathsf{T}} \\ &\text{(c)} \quad (c\mathbf{A})^{\mathsf{T}} = c\mathbf{A}^{\mathsf{T}} \\ &\text{(d)} \quad (\mathbf{AB})^{\mathsf{T}} = \mathbf{B}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}. \end{aligned} \tag{10}$$

CAUTION! Note that in (10d) the transposed matrices are ***in reversed order***. We leave the proofs as an exercise in Probs. 9 and 10.

Special Matrices

Certain kinds of matrices will occur quite frequently in our work, and we now list the most important ones of them.

Symmetric and Skew-Symmetric Matrices. Transposition gives rise to two useful classes of matrices. **Symmetric** matrices are square matrices whose transpose equals the

matrix itself. **Skew-symmetric** matrices are square matrices whose transpose equals ***minus*** the matrix. Both cases are defined in (11) and illustrated by Example 8.

$$\mathbf{A}^\mathsf{T} = \mathbf{A} \quad (\text{thus } a_{kj} = a_{jk}), \qquad \mathbf{A}^\mathsf{T} = -\mathbf{A} \quad (\text{thus } a_{kj} = -a_{jk}, \text{ hence } a_{jj} = 0). \tag{11}$$

Symmetric Matrix — Skew-Symmetric Matrix

EXAMPLE 8 Symmetric and Skew-Symmetric Matrices

$$\mathbf{A} = \begin{bmatrix} 20 & 120 & 200 \\ 120 & 10 & 150 \\ 200 & 150 & 30 \end{bmatrix} \text{ is symmetric, and } \quad \mathbf{B} = \begin{bmatrix} 0 & 1 & -3 \\ -1 & 0 & -2 \\ 3 & 2 & 0 \end{bmatrix} \text{ is skew-symmetric.}$$

For instance, if a company has three building supply centers C_1, C_2, C_3, then **A** could show costs, say, a_{jj} for handling 1000 bags of cement at center C_j, and a_{jk} $(j \neq k)$ the cost of shipping 1000 bags from C_j to C_k. Clearly, $a_{jk} = a_{kj}$ if we assume shipping in the opposite direction will cost the same.

Symmetric matrices have several general properties which make them important. This will be seen as we proceed. ■

Triangular Matrices. **Upper triangular matrices** are square matrices that can have nonzero entries only on and ***above*** the main diagonal, whereas any entry below the diagonal must be zero. Similarly, **lower triangular matrices** can have nonzero entries only on and ***below*** the main diagonal. Any entry on the main diagonal of a triangular matrix may be zero or not.

EXAMPLE 9 Upper and Lower Triangular Matrices

$$\begin{bmatrix} 1 & 3 \\ 0 & 2 \end{bmatrix}, \quad \begin{bmatrix} 1 & 4 & 2 \\ 0 & 3 & 2 \\ 0 & 0 & 6 \end{bmatrix}, \quad \begin{bmatrix} 2 & 0 & 0 \\ 8 & -1 & 0 \\ 7 & 6 & 8 \end{bmatrix}, \quad \begin{bmatrix} 3 & 0 & 0 & 0 \\ 9 & -3 & 0 & 0 \\ 1 & 0 & 2 & 0 \\ 1 & 9 & 3 & 6 \end{bmatrix}.$$ ■

Upper triangular — Lower triangular

Diagonal Matrices. These are square matrices that can have nonzero entries only on the main diagonal. Any entry above or below the main diagonal must be zero.

If all the diagonal entries of a diagonal matrix **S** are equal, say, c, we call **S** a **scalar matrix** because multiplication of any square matrix **A** of the same size by **S** has the same effect as the multiplication by a scalar, that is,

$$\mathbf{AS} = \mathbf{SA} = c\mathbf{A}. \tag{12}$$

In particular, a scalar matrix, whose entries on the main diagonal are all 1, is called a **unit matrix** (or **identity matrix**) and is denoted by $\mathbf{I}_n$ or simply by **I**. For **I**, formula (12) becomes

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}. \tag{13}$$

EXAMPLE 10 Diagonal Matrix D. Scalar Matrix S. Unit Matrix I

$$\mathbf{D} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & -3 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{S} = \begin{bmatrix} c & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & c \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$ ■

Some Applications of Matrix Multiplication

EXAMPLE 11 Computer Production. Matrix Times Matrix

Supercomp Ltd produces two computer models PC1086 and PC1186. The matrix **A** shows the cost per computer (in thousands of dollars) and **B** the production figures for the year 2010 (in multiples of 10,000 units.) Find a matrix **C** that shows the shareholders the cost per quarter (in millions of dollars) for raw material, labor, and miscellaneous.

$$\begin{array}{c} \\ \mathbf{A} = \end{array} \begin{array}{c} \text{PC1086} \quad \text{PC1186} \\ \begin{bmatrix} 1.2 & 1.6 \\ 0.3 & 0.4 \\ 0.5 & 0.6 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array} \qquad \mathbf{B} = \begin{array}{c} \text{Quarter} \\ 1 \quad 2 \quad 3 \quad 4 \\ \begin{bmatrix} 3 & 8 & 6 & 9 \\ 6 & 2 & 4 & 3 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{PC1086} \\ \text{PC1186} \end{array}$$

Solution.

$$\mathbf{C} = \mathbf{AB} = \begin{array}{c} \text{Quarter} \\ 1 \qquad 2 \qquad 3 \qquad 4 \\ \begin{bmatrix} 13.2 & 12.8 & 13.6 & 15.6 \\ 3.3 & 3.2 & 3.4 & 3.9 \\ 5.1 & 5.2 & 5.4 & 6.3 \end{bmatrix} \end{array} \begin{array}{l} \\ \\ \text{Raw Components} \\ \text{Labor} \\ \text{Miscellaneous} \end{array}$$

Since cost is given in multiples of \$1000 and production in multiples of 10,000 units, the entries of **C** are multiples of \$10 millions; thus $c_{11} = 13.2$ means \$132 million, etc. ∎

EXAMPLE 12 Weight Watching. Matrix Times Vector

Suppose that in a weight-watching program, a person of 185 lb burns 350 cal/hr in walking (3 mph), 500 in bicycling (13 mph), and 950 in jogging (5.5 mph). Bill, weighing 185 lb, plans to exercise according to the matrix shown. Verify the calculations (W = Walking, B = Bicycling, J = Jogging).

$$\begin{array}{l} \\ \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array} \begin{array}{c} \text{W} \qquad \text{B} \qquad \text{J} \\ \begin{bmatrix} 1.0 & 0 & 0.5 \\ 1.0 & 1.0 & 0.5 \\ 1.5 & 0 & 0.5 \\ 2.0 & 1.5 & 1.0 \end{bmatrix} \end{array} \begin{bmatrix} 350 \\ 500 \\ 950 \end{bmatrix} = \begin{bmatrix} 825 \\ 1325 \\ 1000 \\ 2400 \end{bmatrix} \begin{array}{l} \text{MON} \\ \text{WED} \\ \text{FRI} \\ \text{SAT} \end{array}$$ ∎

EXAMPLE 13 Markov Process. Powers of a Matrix. Stochastic Matrix

Suppose that the 2004 state of land use in a city of 60 mi^2 of built-up area is

C: Commercially Used 25% I: Industrially Used 20% R: Residentially Used 55%.

Find the states in 2009, 2014, and 2019, assuming that the transition probabilities for 5-year intervals are given by the matrix **A** and remain practically the same over the time considered.

$$\mathbf{A} = \begin{array}{c} \text{From C} \quad \text{From I} \quad \text{From R} \\ \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \end{array} \begin{array}{l} \\ \text{To C} \\ \text{To I} \\ \text{To R} \end{array}$$

A is a **stochastic matrix**, that is, a square matrix with all entries nonnegative and all column sums equal to 1. Our example concerns a **Markov process**,[1] that is, a process for which the probability of entering a certain state depends only on the last state occupied (and the matrix **A**), not on any earlier state.

Solution. From the matrix **A** and the 2004 state we can compute the 2009 state,

$$\begin{matrix} \text{C} \\ \text{I} \\ \text{R} \end{matrix} \begin{bmatrix} 0.7 \cdot 25 + 0.1 \cdot 20 + 0 \quad \cdot 55 \\ 0.2 \cdot 25 + 0.9 \cdot 20 + 0.2 \cdot 55 \\ 0.1 \cdot 25 + \quad 0 \cdot 20 + 0.8 \cdot 55 \end{bmatrix} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix} \begin{bmatrix} 25 \\ 20 \\ 55 \end{bmatrix} = \begin{bmatrix} 19.5 \\ 34.0 \\ 46.5 \end{bmatrix}.$$

To explain: The 2009 figure for C equals 25% times the probability 0.7 that C goes into C, plus 20% times the probability 0.1 that I goes into C, plus 55% times the probability 0 that R goes into C. Together,

$$25 \cdot 0.7 + 20 \cdot 0.1 + 55 \cdot 0 = 19.5\,[\%]. \qquad \text{Also} \qquad 25 \cdot 0.2 + 20 \cdot 0.9 + 55 \cdot 0.2 = 34\,[\%].$$

Similarly, the new R is 46.5%. We see that the 2009 state vector is the column vector

$$\mathbf{y} = [19.5 \quad 34.0 \quad 46.5]^{\mathsf{T}} = \mathbf{Ax} = \mathbf{A}\,[25 \quad 20 \quad 55]^{\mathsf{T}}$$

where the column vector $\mathbf{x} = [25 \quad 20 \quad 55]^{\mathsf{T}}$ is the given 2004 state vector. Note that the sum of the entries of $\mathbf{y}$ is 100 [%]. Similarly, you may verify that for 2014 and 2019 we get the state vectors

$$\mathbf{z} = \mathbf{Ay} = \mathbf{A}(\mathbf{Ax}) = \mathbf{A}^2\mathbf{x} = [17.05 \quad 43.80 \quad 39.15]^{\mathsf{T}}$$
$$\mathbf{u} = \mathbf{Az} = \mathbf{A}^2\mathbf{y} = \mathbf{A}^3\mathbf{x} = [16.315 \quad 50.660 \quad 33.025]^{\mathsf{T}}.$$

Answer. In 2009 the commercial area will be 19.5% (11.7 mi^2), the industrial 34% (20.4 mi^2), and the residential 46.5% (27.9 mi^2). For 2014 the corresponding figures are 17.05%, 43.80%, and 39.15%. For 2019 they are 16.315%, 50.660%, and 33.025%. (In Sec. 8.2 we shall see what happens in the limit, assuming that those probabilities remain the same. In the meantime, can you experiment or guess?) ■

PROBLEM SET 7.2

1–10 GENERAL QUESTIONS

1. **Multiplication.** Why is multiplication of matrices restricted by conditions on the factors?
2. **Square matrix.** What form does a 3 × 3 matrix have if it is symmetric as well as skew-symmetric?
3. **Product of vectors.** Can every 3 × 3 matrix be represented by two vectors as in Example 3?
4. **Skew-symmetric matrix.** How many different entries can a 4 × 4 skew-symmetric matrix have? An $n \times n$ skew-symmetric matrix?
5. Same questions as in Prob. 4 for symmetric matrices.
6. **Triangular matrix.** If $\mathbf{U}_1, \mathbf{U}_2$ are upper triangular and $\mathbf{L}_1, \mathbf{L}_2$ are lower triangular, which of the following are triangular?

$$\mathbf{U}_1 + \mathbf{U}_2, \quad \mathbf{U}_1\mathbf{U}_2, \quad \mathbf{U}_1^2, \quad \mathbf{U}_1 + \mathbf{L}_1, \quad \mathbf{U}_1\mathbf{L}_1, \quad \mathbf{L}_1 + \mathbf{L}_2$$

7. **Idempotent matrix**, defined by $\mathbf{A}^2 = \mathbf{A}$. Can you find four 2 × 2 idempotent matrices?
8. **Nilpotent matrix**, defined by $\mathbf{B}^m = \mathbf{0}$ for some m. Can you find three 2 × 2 nilpotent matrices?
9. **Transposition.** Can you prove (10a)–(10c) for 3 × 3 matrices? For $m \times n$ matrices?
10. **Transposition.** **(a)** Illustrate (10d) by simple examples. **(b)** Prove (10d).

11–20 MULTIPLICATION, ADDITION, AND TRANSPOSITION OF MATRICES AND VECTORS

Let

$$\mathbf{A} = \begin{bmatrix} 4 & -2 & 3 \\ -2 & 1 & 6 \\ 1 & 2 & 2 \end{bmatrix}, \qquad \mathbf{B} = \begin{bmatrix} 1 & -3 & 0 \\ -3 & 1 & 0 \\ 0 & 0 & -2 \end{bmatrix}$$

$$\mathbf{C} = \begin{bmatrix} 0 & 1 \\ 3 & 2 \\ -2 & 0 \end{bmatrix}, \qquad \mathbf{a} = [1 \quad -2 \quad 0], \quad \mathbf{b} = \begin{bmatrix} 3 \\ 1 \\ -1 \end{bmatrix}.$$

[1] ANDREI ANDREJEVITCH MARKOV (1856–1922), Russian mathematician, known for his work in probability theory.

Showing all intermediate results, calculate the following expressions or give reasons why they are undefined:

11. $\mathbf{AB}$, $\mathbf{AB}^\mathsf{T}$, $\mathbf{BA}$, $\mathbf{B}^\mathsf{T}\mathbf{A}$

12. $\mathbf{AA}^\mathsf{T}$, $\mathbf{A}^2$, $\mathbf{BB}^\mathsf{T}$, $\mathbf{B}^2$

13. $\mathbf{CC}^\mathsf{T}$, $\mathbf{BC}$, $\mathbf{CB}$, $\mathbf{C}^\mathsf{T}\mathbf{B}$

14. $3\mathbf{A} - 2\mathbf{B}$, $(3\mathbf{A} - 2\mathbf{B})^\mathsf{T}$, $3\mathbf{A}^\mathsf{T} - 2\mathbf{B}^\mathsf{T}$, $(3\mathbf{A} - 2\mathbf{B})^\mathsf{T}\mathbf{a}^\mathsf{T}$

15. $\mathbf{Aa}$, $\mathbf{Aa}^\mathsf{T}$, $(\mathbf{Ab})^\mathsf{T}$, $\mathbf{b}^\mathsf{T}\mathbf{A}^\mathsf{T}$

16. $\mathbf{BC}$, $\mathbf{BC}^\mathsf{T}$, $\mathbf{Bb}$, $\mathbf{b}^\mathsf{T}\mathbf{B}$

17. $\mathbf{ABC}$, $\mathbf{ABa}$, $\mathbf{ABb}$, $\mathbf{Ca}^\mathsf{T}$

18. $\mathbf{ab}$, $\mathbf{ba}$, $\mathbf{aA}$, $\mathbf{Bb}$

19. $1.5\mathbf{a} + 3.0\mathbf{b}$, $1.5\mathbf{a}^\mathsf{T} + 3.0\mathbf{b}$, $(\mathbf{A} - \mathbf{B})\mathbf{b}$, $\mathbf{Ab} - \mathbf{Bb}$

20. $\mathbf{b}^\mathsf{T}\mathbf{Ab}$, $\mathbf{aBa}^\mathsf{T}$, $\mathbf{aCC}^\mathsf{T}$, $\mathbf{C}^\mathsf{T}\mathbf{ba}$

21. General rules. Prove (2) for 2×2 matrices $\mathbf{A} = [a_{jk}]$, $\mathbf{B} = [b_{jk}]$, $\mathbf{C} = [c_{jk}]$, and a general scalar.

22. Product. Write $\mathbf{AB}$ in Prob. 11 in terms of row and column vectors.

23. Product. Calculate $\mathbf{AB}$ in Prob. 11 columnwise. See Example 1.

24. Commutativity. Find all 2×2 matrices $\mathbf{A} = [a_{jk}]$ that commute with $\mathbf{B} = [b_{jk}]$, where $b_{jk} = j + k$.

25. TEAM PROJECT. Symmetric and Skew-Symmetric Matrices. These matrices occur quite frequently in applications, so it is worthwhile to study some of their most important properties.

(a) Verify the claims in (11) that $a_{kj} = a_{jk}$ for a symmetric matrix, and $a_{kj} = -a_{jk}$ for a skew-symmetric matrix. Give examples.

(b) Show that for every square matrix $\mathbf{C}$ the matrix $\mathbf{C} + \mathbf{C}^\mathsf{T}$ is symmetric and $\mathbf{C} - \mathbf{C}^\mathsf{T}$ is skew-symmetric. Write $\mathbf{C}$ in the form $\mathbf{C} = \mathbf{S} + \mathbf{T}$, where $\mathbf{S}$ is symmetric and $\mathbf{T}$ is skew-symmetric and find $\mathbf{S}$ and $\mathbf{T}$ in terms of $\mathbf{C}$. Represent $\mathbf{A}$ and $\mathbf{B}$ in Probs. 11–20 in this form.

(c) A **linear combination** of matrices $\mathbf{A}, \mathbf{B}, \mathbf{C}, \cdots, \mathbf{M}$ of the same size is an expression of the form

$$a\mathbf{A} + b\mathbf{B} + c\mathbf{C} + \cdots + m\mathbf{M}, \tag{14}$$

where $a, \cdots, m$ are any scalars. Show that if these matrices are square and symmetric, so is (14); similarly, if they are skew-symmetric, so is (14).

(d) Show that $\mathbf{AB}$ with symmetric $\mathbf{A}$ and $\mathbf{B}$ is symmetric if and only if $\mathbf{A}$ and $\mathbf{B}$ **commute**, that is, $\mathbf{AB} = \mathbf{BA}$.

(e) Under what condition is the product of skew-symmetric matrices skew-symmetric?

26–30 FURTHER APPLICATIONS

26. Production. In a production process, let N mean "no trouble" and T "trouble." Let the transition probabilities from one day to the next be 0.8 for $N \to N$, hence 0.2 for $N \to T$, and 0.5 for $T \to N$, hence 0.5 for $T \to T$.

If today there is no trouble, what is the probability of N two days after today? Three days after today?

27. CAS Experiment. Markov Process. Write a program for a Markov process. Use it to calculate further steps in Example 13 of the text. Experiment with other stochastic 3×3 matrices, also using different starting values.

28. Concert subscription. In a community of 100,000 adults, subscribers to a concert series tend to renew their subscription with probability 90% and persons presently not subscribing will subscribe for the next season with probability 0.2%. If the present number of subscribers is 1200, can one predict an increase, decrease, or no change over each of the next three seasons?

29. Profit vector. Two factory outlets F_1 and F_2 in New York and Los Angeles sell sofas (S), chairs (C), and tables (T) with a profit of \$35, \$62, and \$30, respectively. Let the sales in a certain week be given by the matrix

$$\mathbf{A} = \begin{matrix} & \text{S} & \text{C} & \text{T} & \\ & \left[\begin{matrix} 400 \\ 100 \end{matrix}\right. & \begin{matrix} 60 \\ 120 \end{matrix} & \left.\begin{matrix} 240 \\ 500 \end{matrix}\right] & \begin{matrix} F_1 \\ F_2 \end{matrix} \end{matrix}$$

Introduce a "profit vector" $\mathbf{p}$ such that the components of $\mathbf{v} = \mathbf{Ap}$ give the total profits of F_1 and F_2.

30. TEAM PROJECT. Special Linear Transformations. Rotations have various applications. We show in this project how they can be handled by matrices.

(a) Rotation in the plane. Show that the linear transformation $\mathbf{y} = \mathbf{Ax}$ with

$$\mathbf{A} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}$$

is a counterclockwise rotation of the Cartesian x_1x_2-coordinate system in the plane about the origin, where θ is the angle of rotation.

(b) Rotation through $n\theta$. Show that in (a)

$$\mathbf{A}^n = \begin{bmatrix} \cos n\theta & -\sin n\theta \\ \sin n\theta & \cos n\theta \end{bmatrix}.$$

Is this plausible? Explain this in words.

(c) Addition formulas for cosine and sine. By geometry we should have

$$\begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix} \begin{bmatrix} \cos\beta & -\sin\beta \\ \sin\beta & \cos\beta \end{bmatrix} = \begin{bmatrix} \cos(\alpha+\beta) & -\sin(\alpha+\beta) \\ \sin(\alpha+\beta) & \cos(\alpha+\beta) \end{bmatrix}.$$

Derive from this the addition formulas (6) in App. A3.1.

(d) Computer graphics. To visualize a three-dimensional object with plane faces (e.g., a cube), we may store the position vectors of the vertices with respect to a suitable $x_1x_2x_3$-coordinate system (and a list of the connecting edges) and then obtain a two-dimensional image on a video screen by projecting the object onto a coordinate plane, for instance, onto the x_1x_2-plane by setting $x_3 = 0$. To change the appearance of the image, we can impose a linear transformation on the position vectors stored. Show that a diagonal matrix **D** with main diagonal entries 3, 1, $\frac{1}{2}$ gives from an $\mathbf{x} = [x_j]$ the new position vector $\mathbf{y} = \mathbf{D}\mathbf{x}$, where $y_1 = 3x_1$ (stretch in the x_1-direction by a factor 3), $y_2 = x_2$ (unchanged), $y_3 = \frac{1}{2}x_3$ (contraction in the x_3-direction). What effect would a scalar matrix have?

(e) Rotations in space. Explain $\mathbf{y} = \mathbf{A}\mathbf{x}$ geometrically when **A** is one of the three matrices

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix},$$

$$\begin{bmatrix} \cos\varphi & 0 & -\sin\varphi \\ 0 & 1 & 0 \\ \sin\varphi & 0 & \cos\varphi \end{bmatrix}, \quad \begin{bmatrix} \cos\psi & -\sin\psi & 0 \\ \sin\psi & \cos\psi & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

What effect would these transformations have in situations such as that described in (d)?

7.3 Linear Systems of Equations. Gauss Elimination

We now come to one of the most important use of matrices, that is, using matrices to solve systems of linear equations. We showed informally, in Example 1 of Sec. 7.1, how to represent the information contained in a system of linear equations by a matrix, called the augmented matrix. This matrix will then be used in solving the linear system of equations. Our approach to solving linear systems is called the Gauss elimination method. Since this method is so fundamental to linear algebra, the student should be alert.

A shorter term for systems of linear equations is just **linear systems**. Linear systems model many applications in engineering, economics, statistics, and many other areas. Electrical networks, traffic flow, and commodity markets may serve as specific examples of applications.

Linear System, Coefficient Matrix, Augmented Matrix

A **linear system of *m* equations in *n* unknowns** $x_1, \cdots, x_n$ is a set of equations of the form

(1)
$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots &\cdots\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m. \end{aligned}$$

The system is called *linear* because each variable x_j appears in the first power only, just as in the equation of a straight line. $a_{11}, \cdots, a_{mn}$ are given numbers, called the **coefficients** of the system. $b_1, \cdots, b_m$ on the right are also given numbers. If all the b_j are zero, then (1) is called a **homogeneous system**. If at least one b_j is not zero, then (1) is called a **nonhomogeneous system**.

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfies all the m equations. A **solution vector** of (1) is a vector $\mathbf{x}$ whose components form a solution of (1). If the system (1) is homogeneous, it always has at least the **trivial solution** $x_1 = 0, \cdots, x_n = 0$.

Matrix Form of the Linear System (1). From the definition of matrix multiplication we see that the m equations of (1) may be written as a single vector equation

$$\mathbf{Ax} = \mathbf{b} \tag{2}$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $m \times n$ matrix

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \cdot \\ \cdot \\ \cdot \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_m \end{bmatrix}$$

are column vectors. We assume that the coefficients a_{jk} are not all zero, so that $\mathbf{A}$ is not a zero matrix. Note that $\mathbf{x}$ has n components, whereas $\mathbf{b}$ has m components. The matrix

$$\tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right]$$

is called the **augmented matrix** of the system (1). The dashed vertical line could be omitted, as we shall do later. It is merely a reminder that the last column of $\tilde{\mathbf{A}}$ did not come from matrix $\mathbf{A}$ but came from vector $\mathbf{b}$. Thus, we *augmented* the matrix $\mathbf{A}$.

*Note that **the augmented matrix** $\tilde{\mathbf{A}}$ **determines the system (1) completely*** because it contains all the given numbers appearing in (1).

EXAMPLE 1 Geometric Interpretation. Existence and Uniqueness of Solutions

If $m = n = 2$, we have two equations in two unknowns x_1, x_2

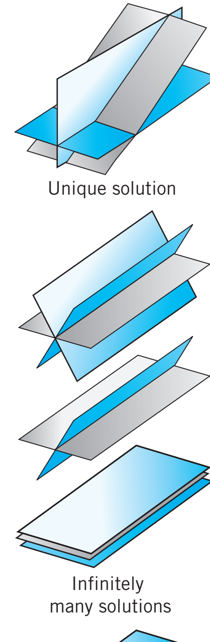
$$a_{11}x_1 + a_{12}x_2 = b_1$$
$$a_{21}x_1 + a_{22}x_2 = b_2.$$

If we interpret x_1, x_2 as coordinates in the x_1x_2-plane, then each of the two equations represents a straight line, and (x_1, x_2) is a solution if and only if the point P with coordinates x_1, x_2 lies on both lines. Hence there are three possible cases (see Fig. 158 on next page):

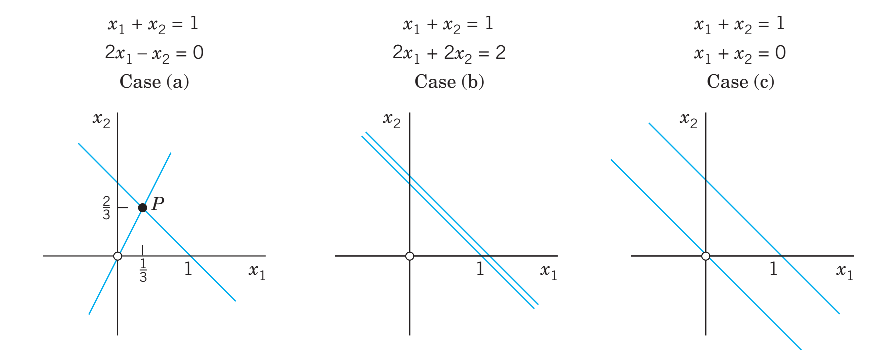
(*a*) Precisely one solution if the lines intersect

(*b*) Infinitely many solutions if the lines coincide

(*c*) No solution if the lines are parallel



For instance,



If the system is homogenous, Case (c) cannot happen, because then those two straight lines pass through the origin, whose coordinates (0, 0) constitute the trivial solution. Similarly, our present discussion can be extended from two equations in two unknowns to three equations in three unknowns. We give the geometric interpretation of three possible cases concerning solutions in Fig. 158. Instead of straight lines we have planes and the solution depends on the positioning of these planes in space relative to each other. The student may wish to come up with some specific examples. ■

Our simple example illustrated that a system (1) may have no solution. This leads to such questions as: Does a given system (1) have a solution? Under what conditions does it have precisely one solution? If it has more than one solution, how can we characterize the set of all solutions? We shall consider such questions in Sec. 7.5.

First, however, let us discuss an important systematic method for solving linear systems.

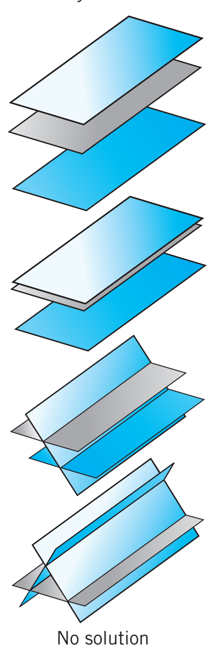


Fig. 158. Three equations in three unknowns interpreted as planes in space

Gauss Elimination and Back Substitution

The Gauss elimination method can be motivated as follows. Consider a linear system that is in ***triangular form*** (in full, *upper* triangular form) such as

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ 13x_2 &= -26 \end{aligned}$$

(*Triangular* means that all the nonzero entries of the corresponding coefficient matrix lie above the diagonal and form an upside-down 90° triangle.) Then we can solve the system by **back substitution**, that is, we solve the last equation for the variable, $x_2 = -26/13 = -2$, and then work backward, substituting $x_2 = -2$ into the first equation and solving it for x_1, obtaining $x_1 = \frac{1}{2}(2 - 5x_2) = \frac{1}{2}(2 - 5 \cdot (-2)) = 6$. This gives us the idea of first reducing a general system to triangular form. For instance, let the given system be

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ -4x_1 + 3x_2 &= -30. \end{aligned} \qquad \text{Its augmented matrix is} \qquad \begin{bmatrix} 2 & 5 & 2 \\ -4 & 3 & -30 \end{bmatrix}.$$

We leave the first equation as it is. We eliminate x_1 from the second equation, to get a triangular system. For this we add twice the first equation to the second, and we do the same

operation on the ***rows*** of the augmented matrix. This gives $-4x_1 + 4x_1 + 3x_2 + 10x_2 = -30 + 2 \cdot 2$, that is,

$$\begin{aligned} 2x_1 + 5x_2 &= 2 \\ 13x_2 &= -26 \end{aligned} \qquad \text{Row 2} + 2\ \text{Row 1} \qquad \begin{bmatrix} 2 & 5 & 2 \\ 0 & 13 & -26 \end{bmatrix}$$

where Row 2 + 2 Row 1 means "Add twice Row 1 to Row 2" in the original matrix. This is the **Gauss elimination** (for 2 equations in 2 unknowns) giving the triangular form, from which back substitution now yields $x_2 = -2$ and $x_1 = 6$, as before.

Since a linear system is completely determined by its augmented matrix, ***Gauss elimination can be done by merely considering the matrices***, as we have just indicated. We do this again in the next example, emphasizing the matrices by writing them first and the equations behind them, just as a help in order not to lose track.

EXAMPLE 2 Gauss Elimination. Electrical Network

Solve the linear system

$$\begin{aligned} x_1 - x_2 + x_3 &= 0 \\ -x_1 + x_2 - x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ 20x_1 + 10x_2 &= 80. \end{aligned}$$

Derivation from the circuit in Fig. 159 (Optional). This is the system for the unknown currents $x_1 = i_1$, $x_2 = i_2$, $x_3 = i_3$ in the electrical network in Fig. 159. To obtain it, we label the currents as shown, choosing directions arbitrarily; if a current will come out negative, this will simply mean that the current flows against the direction of our arrow. The current entering each battery will be the same as the current leaving it. The equations for the currents result from Kirchhoff's laws:

Kirchhoff's Current Law (KCL). *At any point of a circuit, the sum of the inflowing currents equals the sum of the outflowing currents.*

Kirchhoff's Voltage Law (KVL). *In any closed loop, the sum of all voltage drops equals the impressed electromotive force.*

Node P gives the first equation, node Q the second, the right loop the third, and the left loop the fourth, as indicated in the figure.

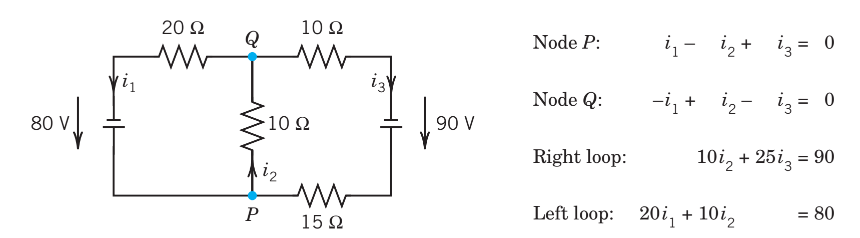
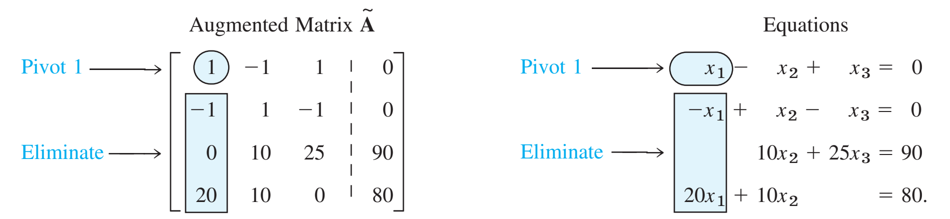


Fig. 159. Network in Example 2 and equations relating the currents

Solution by Gauss Elimination. This system could be solved rather quickly by noticing its particular form. But this is not the point. The point is that the Gauss elimination is systematic and will work in general,

also for large systems. We apply it to our system and then do back substitution. As indicated, let us write the augmented matrix of the system first and then the system itself:



Step 1. Elimination of x_1

Call the first row of $\mathbf{A}$ the **pivot row** and the first equation the **pivot equation**. Call the coefficient 1 of its x_1-term the **pivot** in this step. Use this equation to eliminate x_1 (get rid of x_1) in the other equations. For this, do:

Add 1 times the pivot equation to the second equation.

Add -20 times the pivot equation to the fourth equation.

This corresponds to **row operations** on the augmented matrix as indicated in BLUE behind the ***new matrix*** in (3). So the operations are performed on the ***preceding matrix***. The result is

$$(3)\qquad \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 30 & -20 & 80 \end{array}\right] \quad \begin{array}{l} \\ \text{Row 2 + Row 1} \\ \\ \text{Row 4 − 20 Row 1} \end{array} \quad \begin{aligned} x_1 - x_2 + x_3 &= 0 \\ 0 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ 30x_2 - 20x_3 &= 80. \end{aligned}$$

Step 2. Elimination of x_2

The first equation remains as it is. We want the new second equation to serve as the next pivot equation. But since it has no x_2-term (in fact, it is $0 = 0$), we must first change the order of the equations and the corresponding rows of the new matrix. We put $0 = 0$ at the end and move the third equation and the fourth equation one place up. This is called **partial pivoting** (as opposed to the rarely used *total pivoting*, in which the order of the unknowns is also changed). It gives

$$\begin{array}{l} \\ \text{Pivot 10} \longrightarrow \\ \text{Eliminate 30} \longrightarrow \\ \\ \end{array} \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & \boxed{10} & 25 & 90 \\ 0 & \boxed{30} & -20 & 80 \\ 0 & 0 & 0 & 0 \end{array}\right] \qquad \begin{array}{l} \\ \text{Pivot 10} \longrightarrow \\ \text{Eliminate } 30x_2 \longrightarrow \\ \\ \end{array} \begin{aligned} x_1 - x_2 + x_3 &= 0 \\ \boxed{10x_2} + 25x_3 &= 90 \\ \boxed{30x_2} - 20x_3 &= 80 \\ 0 &= 0. \end{aligned}$$

To eliminate x_2, do:

Add -3 times the pivot equation to the third equation.

The result is

$$(4)\qquad \left[\begin{array}{rrr|r} 1 & -1 & 1 & 0 \\ 0 & 10 & 25 & 90 \\ 0 & 0 & -95 & -190 \\ 0 & 0 & 0 & 0 \end{array}\right] \quad \begin{array}{l} \\ \\ \text{Row 3 − 3 Row 2} \\ \\ \end{array} \quad \begin{aligned} x_1 - x_2 + x_3 &= 0 \\ 10x_2 + 25x_3 &= 90 \\ -95x_3 &= -190 \\ 0 &= 0. \end{aligned}$$

Back Substitution. ***Determination of x_3, x_2, x_1 (in this order)***

Working backward from the last to the first equation of this "triangular" system (4), we can now readily find x_3, then x_2, and then x_1:

$$\begin{aligned} -95x_3 &= -190 \qquad & x_3 &= i_3 = 2\ [\text{A}] \\ 10x_2 + 25x_3 &= 90 \qquad & x_2 &= \tfrac{1}{10}(90 - 25x_3) = i_2 = 4\ [\text{A}] \\ x_1 - x_2 + x_3 &= 0 \qquad & x_1 &= x_2 - x_3 = i_1 = 2\ [\text{A}] \end{aligned}$$

where A stands for "amperes." This is the answer to our problem. The solution is unique. ■

Elementary Row Operations. Row-Equivalent Systems

Example 2 illustrates the operations of the Gauss elimination. These are the first two of three operations, which are called

Elementary Row Operations for Matrices:

Interchange of two rows

Addition of a constant multiple of one row to another row

Multiplication of a row by a ***nonzero*** *constant c*

CAUTION! These operations are for rows, ***not for columns***! They correspond to the following

Elementary Operations for Equations:

Interchange of two equations

Addition of a constant multiple of one equation to another equation

Multiplication of an equation by a ***nonzero*** *constant c*

Clearly, the interchange of two equations does not alter the solution set. Neither does their addition because we can undo it by a corresponding subtraction. Similarly for their multiplication, which we can undo by multiplying the new equation by $1/c$ (since $c \neq 0$), producing the original equation.

We now call a linear system S_1 **row-equivalent** to a linear system S_2 if S_1 can be obtained from S_2 by (finitely many!) row operations. This justifies Gauss elimination and establishes the following result.

THEOREM 1

Row-Equivalent Systems

Row-equivalent linear systems have the same set of solutions.

Because of this theorem, systems having the same solution sets are often called *equivalent systems*. But note well that we are dealing with ***row operations***. No column operations on the augmented matrix are permitted in this context because they would generally alter the solution set.

A linear system (1) is called **overdetermined** if it has more equations than unknowns, as in Example 2, **determined** if $m = n$, as in Example 1, and **underdetermined** if it has fewer equations than unknowns.

Furthermore, a system (1) is called **consistent** if it has at least one solution (thus, one solution or infinitely many solutions), but **inconsistent** if it has no solutions at all, as $x_1 + x_2 = 1, x_1 + x_2 = 0$ in Example 1, Case (c).

Gauss Elimination: The Three Possible Cases of Systems

We have seen, in Example 2, that Gauss elimination can solve linear systems that have a unique solution. This leaves us to apply Gauss elimination to a system with infinitely many solutions (in Example 3) and one with no solution (in Example 4).

EXAMPLE 3 Gauss Elimination if Infinitely Many Solutions Exist

Solve the following linear system of three equations in four unknowns whose augmented matrix is

$$(5)\quad \begin{bmatrix} 3.0 & 2.0 & 2.0 & -5.0 & | & 8.0 \\ 0.6 & 1.5 & 1.5 & -5.4 & | & 2.7 \\ 1.2 & -0.3 & -0.3 & 2.4 & | & 2.1 \end{bmatrix}. \quad \text{Thus,} \quad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 0.6x_1 + 1.5x_2 + 1.5x_3 - 5.4x_4 &= 2.7 \\ 1.2x_1 - 0.3x_2 - 0.3x_3 + 2.4x_4 &= 2.1. \end{aligned}$$

Solution. As in the previous example, we circle pivots and box terms of equations and corresponding entries to be eliminated. We indicate the operations in terms of equations and operate on both equations and matrices.

Step 1. Elimination of x_1 from the second and third equations by adding

$$-0.6/3.0 = -0.2 \text{ times the first equation to the second equation,}$$

$$-1.2/3.0 = -0.4 \text{ times the first equation to the third equation.}$$

This gives the following, in which the pivot of the next step is circled.

$$(6)\quad \begin{bmatrix} 3.0 & 2.0 & 2.0 & -5.0 & | & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & | & 1.1 \\ 0 & -1.1 & -1.1 & 4.4 & | & -1.1 \end{bmatrix} \quad \begin{matrix} \\ \text{Row 2} - 0.2 \text{ Row 1} \\ \text{Row 3} - 0.4 \text{ Row 1} \end{matrix} \quad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 1.1x_2 + 1.1x_3 - 4.4x_4 &= 1.1 \\ -1.1x_2 - 1.1x_3 + 4.4x_4 &= -1.1. \end{aligned}$$

Step 2. Elimination of x_2 from the third equation of (6) by adding

$$1.1/1.1 = 1 \text{ times the second equation to the third equation.}$$

This gives

$$(7)\quad \begin{bmatrix} 3.0 & 2.0 & 2.0 & -5.0 & | & 8.0 \\ 0 & 1.1 & 1.1 & -4.4 & | & 1.1 \\ 0 & 0 & 0 & 0 & | & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ \text{Row 3} + \text{Row 2} \end{matrix} \quad \begin{aligned} 3.0x_1 + 2.0x_2 + 2.0x_3 - 5.0x_4 &= 8.0 \\ 1.1x_2 + 1.1x_3 - 4.4x_4 &= 1.1 \\ 0 &= 0. \end{aligned}$$

Back Substitution. From the second equation, $x_2 = 1 - x_3 + 4x_4$. From this and the first equation, $x_1 = 2 - x_4$. Since x_3 and x_4 remain arbitrary, we have infinitely many solutions. If we choose a value of x_3 and a value of x_4, then the corresponding values of x_1 and x_2 are uniquely determined.

On Notation. If unknowns remain arbitrary, it is also customary to denote them by other letters $t_1, t_2, \cdots$. In this example we may thus write $x_1 = 2 - x_4 = 2 - t_2$, $x_2 = 1 - x_3 + 4x_4 = 1 - t_1 + 4t_2$, $x_3 = t_1$ (first arbitrary unknown), $x_4 = t_2$ (second arbitrary unknown). ■

EXAMPLE 4 Gauss Elimination if no Solution Exists

What will happen if we apply the Gauss elimination to a linear system that has no solution? The answer is that in this case the method will show this fact by producing a contradiction. For instance, consider

$$\begin{bmatrix} 3 & 2 & 1 & | & 3 \\ 2 & 1 & 1 & | & 0 \\ 6 & 2 & 4 & | & 6 \end{bmatrix} \qquad \begin{aligned} 3x_1 + 2x_2 + x_3 &= 3 \\ 2x_1 + x_2 + x_3 &= 0 \\ 6x_1 + 2x_2 + 4x_3 &= 6. \end{aligned}$$

Step 1. Elimination of x_1 from the second and third equations by adding

$$-\tfrac{2}{3} \text{ times the first equation to the second equation,}$$

$$-\tfrac{6}{3} = -2 \text{ times the first equation to the third equation.}$$

This gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & -2 & 2 & 0 \end{array}\right] \quad \begin{array}{l} \\ \text{Row 2} - \frac{2}{3}\,\text{Row 1} \\ \text{Row 3} - 2\,\text{Row 1} \end{array} \qquad \begin{aligned} 3x_1 + 2x_2 + \ x_3 &= \ \ 3 \\ -\tfrac{1}{3}x_2 + \tfrac{1}{3}x_3 &= -2 \\ -\ 2x_2 + 2x_3 &= \ \ 0. \end{aligned}$$

Step 2. Elimination of x_2 from the third equation gives

$$\left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right] \quad \begin{array}{l} \\ \\ \text{Row 3} - 6\,\text{Row 2} \end{array} \qquad \begin{aligned} 3x_1 + 2x_2 + \ x_3 &= \ \ 3 \\ -\tfrac{1}{3}x_2 + \tfrac{1}{3}x_3 &= -\ 2 \\ 0 &= \ 12. \end{aligned}$$

The false statement $0 = 12$ shows that the system has no solution. ■

Row Echelon Form and Information From It

At the end of the Gauss elimination the form of the coefficient matrix, the augmented matrix, and the system itself are called the **row echelon form**. In it, rows of zeros, if present, are the last rows, and, in each nonzero row, the leftmost nonzero entry is farther to the right than in the previous row. For instance, in Example 4 the coefficient matrix and its augmented in row echelon form are

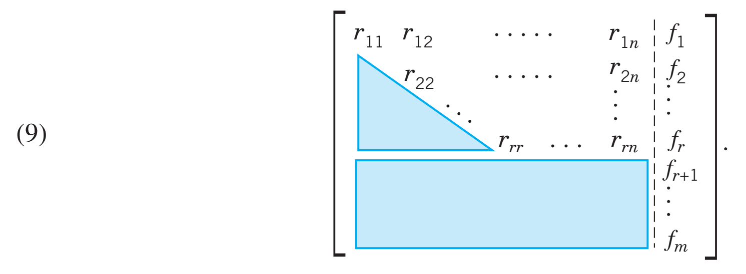
$$\text{(8)} \qquad \begin{bmatrix} 3 & 2 & 1 \\ 0 & -\frac{1}{3} & \frac{1}{3} \\ 0 & 0 & 0 \end{bmatrix} \quad \text{and} \quad \left[\begin{array}{ccc|c} 3 & 2 & 1 & 3 \\ 0 & -\frac{1}{3} & \frac{1}{3} & -2 \\ 0 & 0 & 0 & 12 \end{array}\right].$$

Note that we do not require that the leftmost nonzero entries be 1 since this would have no theoretic or numeric advantage. (The so-called *reduced echelon form*, in which those entries *are* 1, will be discussed in Sec. 7.8.)

The original system of m equations in n unknowns has augmented matrix $[\mathbf{A} \mid \mathbf{b}]$. This is to be row reduced to matrix $[\mathbf{R} \mid \mathbf{f}]$. The two systems $\mathbf{Ax} = \mathbf{b}$ and $\mathbf{Rx} = \mathbf{f}$ are equivalent: if either one has a solution, so does the other, and the solutions are identical.

At the end of the Gauss elimination (before the back substitution), the row echelon form of the augmented matrix will be

(9)



Here, $r \leqq m$, $r_{11} \neq 0$, and all entries in the blue triangle and blue rectangle are zero.

The number of nonzero rows, r, in the row-reduced coefficient matrix $\mathbf{R}$ is called the **rank of R** and also the **rank of A**. Here is the method for determining whether $\mathbf{Ax} = \mathbf{b}$ has solutions and what they are:

(a) No solution. If r is less than m (meaning that $\mathbf{R}$ actually has at least one row of all 0s) *and* at least one of the numbers $f_{r+1}, f_{r+2}, \cdots, f_m$ is not zero, then the system

$\mathbf{Rx} = \mathbf{f}$ is inconsistent: No solution is possible. Therefore the system $\mathbf{Ax} = \mathbf{b}$ is inconsistent as well. See Example 4, where $r = 2 < m = 3$ and $f_{r+1} = f_3 = 12$.

If the system is consistent (either $r = m$, or $r < m$ *and* all the numbers $f_{r+1}, f_{r+2}, \cdots, f_m$ are zero), then there are solutions.

(b) Unique solution. If the system is consistent and $r = n$, there is exactly one solution, which can be found by back substitution. See Example 2, where $r = n = 3$ and $m = 4$.

(c) Infinitely many solutions. To obtain any of these solutions, choose values of $x_{r+1}, \cdots, x_n$ arbitrarily. Then solve the rth equation for x_r (in terms of those arbitrary values), then the $(r - 1)$st equation for x_{r-1}, and so on up the line. See Example 3.

Orientation. Gauss elimination is reasonable in computing time and storage demand. We shall consider those aspects in Sec. 20.1 in the chapter on numeric linear algebra. Section 7.4 develops fundamental concepts of linear algebra such as linear independence and rank of a matrix. These in turn will be used in Sec. 7.5 to fully characterize the behavior of linear systems in terms of existence and uniqueness of solutions.

PROBLEM SET 7.3

1–14 GAUSS ELIMINATION

Solve the linear system given explicitly or by its augmented matrix. Show details.

1. $\begin{aligned} 4x - 6y &= -11 \\ -3x + 8y &= 10 \end{aligned}$

2. $\begin{bmatrix} 3.0 & -0.5 & 0.6 \\ 1.5 & 4.5 & 6.0 \end{bmatrix}$

3. $\begin{aligned} x + y - z &= 9 \\ 8y + 6z &= -6 \\ -2x + 4y - 6z &= 40 \end{aligned}$

4. $\begin{bmatrix} 4 & 1 & 0 & 4 \\ 5 & -3 & 1 & 2 \\ -9 & 2 & -1 & 5 \end{bmatrix}$

5. $\begin{bmatrix} 13 & 12 & -6 \\ -4 & 7 & -73 \\ 11 & -13 & 157 \end{bmatrix}$

6. $\begin{bmatrix} 4 & -8 & 3 & 16 \\ -1 & 2 & -5 & -21 \\ 3 & -6 & 1 & 7 \end{bmatrix}$

7. $\begin{bmatrix} 2 & 4 & 1 & 0 \\ -1 & 1 & -2 & 0 \\ 4 & 0 & 6 & 0 \end{bmatrix}$

8. $\begin{aligned} 4y + 3z &= 8 \\ 2x - z &= 2 \\ 3x + 2y &= 5 \end{aligned}$

9. $\begin{aligned} -2y - 2z &= -8 \\ 3x + 4y - 5z &= 13 \end{aligned}$

10. $\begin{bmatrix} 5 & -7 & 3 & 17 \\ -15 & 21 & -9 & 50 \end{bmatrix}$

11. $\begin{bmatrix} 0 & 5 & 5 & -10 & 0 \\ 2 & -3 & -3 & 6 & 2 \\ 4 & 1 & 1 & -2 & 4 \end{bmatrix}$

12. $\begin{bmatrix} 2 & -2 & 4 & 0 & 0 \\ -3 & 3 & -6 & 5 & 15 \\ 1 & -1 & 2 & 0 & 0 \end{bmatrix}$

13. $\begin{aligned} 10x + 4y - 2z &= -4 \\ -3w - 17x + y + 2z &= 2 \\ w + x + y &= 6 \\ 8w - 34x + 16y - 10z &= 4 \end{aligned}$

14. $\begin{bmatrix} 2 & 3 & 1 & -11 & 1 \\ 5 & -2 & 5 & -4 & 5 \\ 1 & -1 & 3 & -3 & 3 \\ 3 & 4 & -7 & 2 & -7 \end{bmatrix}$

15. Equivalence relation. By definition, an *equivalence relation* on a set is a relation satisfying three conditions: (named as indicated)

(i) Each element A of the set is equivalent to itself (*Reflexivity*).

(ii) If A is equivalent to B, then B is equivalent to A (*Symmetry*).

(iii) If A is equivalent to B and B is equivalent to C, then A is equivalent to C (*Transitivity*).

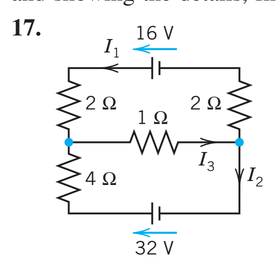
Show that row equivalence of matrices satisfies these three conditions. *Hint.* Show that for each of the three elementary row operations these conditions hold.

16. CAS PROJECT. Gauss Elimination and Back Substitution. Write a program for Gauss elimination and back substitution **(a)** that does not include pivoting and **(b)** that does include pivoting. Apply the programs to Probs. 11–14 and to some larger systems of your choice.

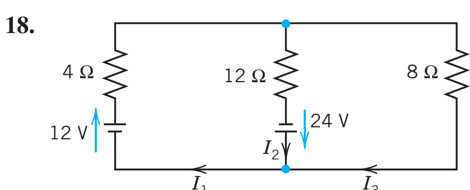
17–21 MODELS OF NETWORKS

In Probs. 17–19, using Kirchhoff's laws (see Example 2) and showing the details, find the currents:

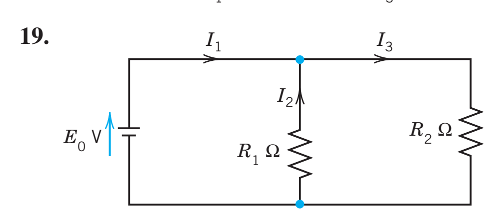
17.



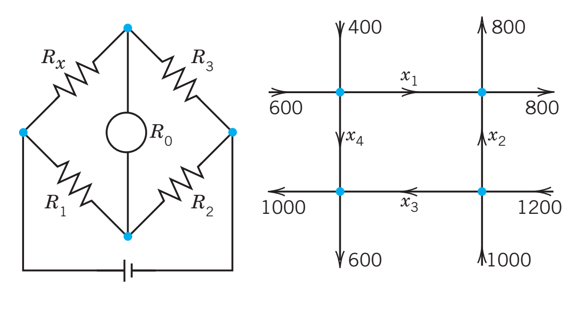
18.



19.



20. Wheatstone bridge. Show that if $R_x/R_3 = R_1/R_2$ in the figure, then $I = 0$. (R_0 is the resistance of the instrument by which I is measured.) This bridge is a method for determining R_x. R_1, R_2, R_3 are known. R_3 is variable. To get R_x, make $I = 0$ by varying R_3. Then calculate $R_x = R_3R_1/R_2$.



Problem 20 Problem 21

21. Traffic flow. Methods of electrical circuit analysis have applications to other fields. For instance, applying the analog of Kirchhoff's Current Law, find the traffic flow (cars per hour) in the net of one-way streets (in the directions indicated by the arrows) shown in the figure. Is the solution unique?

22. Models of markets. Determine the equilibrium solution ($D_1 = S_1, D_2 = S_2$) of the two-commodity market with linear model (D, S, P = demand, supply, price; index 1 = first commodity, index 2 = second commodity)

$$D_1 = 40 - 2P_1 - P_2, \qquad S_1 = 4P_1 - P_2 + 4,$$

$$D_2 = 5P_1 - 2P_2 + 16, \qquad S_2 = 3P_2 - 4.$$

23. Balancing a chemical equation $x_1C_3H_8 + x_2O_2 \rightarrow x_3CO_2 + x_4H_2O$ means finding integer x_1, x_2, x_3, x_4 such that the numbers of atoms of carbon (C), hydrogen (H), and oxygen (O) are the same on both sides of this reaction, in which propane C_3H_8 and O_2 give carbon dioxide and water. Find the smallest positive integers $x_1, \cdots, x_4$.

24. PROJECT. Elementary Matrices. The idea is that elementary operations can be accomplished by matrix multiplication. If **A** is an $m \times n$ matrix on which we want to do an elementary operation, then there is a matrix **E** such that **EA** is the new matrix after the operation. Such an **E** is called an **elementary matrix**. This idea can be helpful, for instance, in the design of algorithms. (*Computationally*, it is generally preferable to do row operations *directly*, rather than by multiplication by **E**.)

(a) Show that the following are elementary matrices, for interchanging Rows 2 and 3, for adding −5 times the first row to the third, and for multiplying the fourth row by 8.

$$\mathbf{E}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{E}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ -5 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

$$\mathbf{E}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}.$$

Apply $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ to a vector and to a 4×3 matrix of your choice. Find $\mathbf{B} = \mathbf{E}_3\mathbf{E}_2\mathbf{E}_1\mathbf{A}$, where $\mathbf{A} = [a_{jk}]$ is the general 4×2 matrix. Is $\mathbf{B}$ equal to $\mathbf{C} = \mathbf{E}_1\mathbf{E}_2\mathbf{E}_3\mathbf{A}$?

(b) Conclude that $\mathbf{E}_1$, $\mathbf{E}_2$, $\mathbf{E}_3$ are obtained by doing the corresponding elementary operations on the 4×4 unit matrix. Prove that *if* $\mathbf{M}$ *is obtained from* $\mathbf{A}$ *by an elementary row operation, then*

$$\mathbf{M} = \mathbf{EA},$$

where $\mathbf{E}$ *is obtained from the* $n \times n$ *unit matrix* $\mathbf{I}_n$ *by the same row operation.*

7.4 Linear Independence. Rank of a Matrix. Vector Space

Since our next goal is to fully characterize the behavior of linear systems in terms of existence and uniqueness of solutions (Sec. 7.5), we have to introduce new fundamental linear algebraic concepts that will aid us in doing so. Foremost among these are ***linear independence*** and the ***rank of a matrix***. Keep in mind that these concepts are intimately linked with the important Gauss elimination method and how it works.

Linear Independence and Dependence of Vectors

Given any set of m vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ (with the same number of components), a **linear combination** of these vectors is an expression of the form

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)}$$

where $c_1, c_2, \cdots, c_m$ are any scalars. Now consider the equation

$$c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0}. \tag{1}$$

Clearly, this vector equation (1) holds if we choose all c_j's zero, because then it becomes $\mathbf{0} = \mathbf{0}$. If this is the only m-tuple of scalars for which (1) holds, then our vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ are said to form a *linearly independent set* or, more briefly, we call them **linearly independent**. Otherwise, if (1) also holds with scalars not all zero, we call these vectors **linearly dependent**. This means that we can express at least one of the vectors as a linear combination of the other vectors. For instance, if (1) holds with, say, $c_1 \neq 0$, we can solve (1) for $\mathbf{a}_{(1)}$:

$$\mathbf{a}_{(1)} = k_2\mathbf{a}_{(2)} + \cdots + k_m\mathbf{a}_{(m)} \qquad \text{where } k_j = -c_j/c_1.$$

(Some k_j's may be zero. Or even all of them, namely, if $\mathbf{a}_{(1)} = \mathbf{0}$.)

Why is linear independence important? Well, if a set of vectors is linearly dependent, then we can get rid of at least one or perhaps more of the vectors until we get a linearly independent set. This set is then the smallest "truly essential" set with which we can work. Thus, we cannot express any of the vectors, of this set, linearly in terms of the others.

EXAMPLE 1 **Linear Independence and Dependence**

The three vectors

$$\begin{aligned} \mathbf{a}_{(1)} &= [\;3 \quad 0 \quad 2 \quad 2\;] \\ \mathbf{a}_{(2)} &= [-6 \quad 42 \quad 24 \quad 54] \\ \mathbf{a}_{(3)} &= [\;21 \quad -21 \quad 0 \quad -15] \end{aligned}$$

are linearly dependent because

$$6\mathbf{a}_{(1)} - \tfrac{1}{2}\mathbf{a}_{(2)} - \mathbf{a}_{(3)} = \mathbf{0}.$$

Although this is easily checked by vector arithmetic (do it!), it is not so easy to discover. However, a systematic method for finding out about linear independence and dependence follows below.

The first two of the three vectors are linearly independent because $c_1\mathbf{a}_{(1)} + c_2\mathbf{a}_{(2)} = \mathbf{0}$ implies $c_2 = 0$ (from the second components) and then $c_1 = 0$ (from any other component of $\mathbf{a}_{(1)}$. ■

Rank of a Matrix

DEFINITION

The **rank** of a matrix $\mathbf{A}$ is the maximum number of linearly independent row vectors of $\mathbf{A}$. It is denoted by rank $\mathbf{A}$.

Our further discussion will show that the rank of a matrix is an important key concept for understanding general properties of matrices and linear systems of equations.

EXAMPLE 2 **Rank**

The matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \tag{2}$$

has rank 2, because Example 1 shows that the first two row vectors are linearly independent, whereas all three row vectors are linearly dependent.

Note further that rank $\mathbf{A} = 0$ if and only if $\mathbf{A} = \mathbf{0}$. This follows directly from the definition. ■

We call a matrix $\mathbf{A}_1$ **row-equivalent** to a matrix $\mathbf{A}_2$ if $\mathbf{A}_1$ can be obtained from $\mathbf{A}_2$ by (finitely many!) elementary row operations.

Now the maximum number of linearly independent row vectors of a matrix does not change if we change the order of rows or multiply a row by a nonzero c or take a linear combination by adding a multiple of a row to another row. This shows that rank is **invariant** under elementary row operations:

THEOREM 1

Row-Equivalent Matrices

Row-equivalent matrices have the same rank.

Hence we can determine the rank of a matrix by reducing the matrix to row-echelon form, as was done in Sec. 7.3. Once the matrix is in row-echelon form, we count the number of nonzero rows, which is precisely the rank of the matrix.

EXAMPLE 3 **Determination of Rank**

For the matrix in Example 2 we obtain successively

$$\mathbf{A} = \begin{bmatrix} 3 & 0 & 2 & 2 \\ -6 & 42 & 24 & 54 \\ 21 & -21 & 0 & -15 \end{bmatrix} \quad \text{(given)}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & -21 & -14 & -29 \end{bmatrix} \quad \begin{matrix} \\ \text{Row 2} + 2 \text{ Row 1} \\ \text{Row 3} - 7 \text{ Row 1} \end{matrix}$$

$$\begin{bmatrix} 3 & 0 & 2 & 2 \\ 0 & 42 & 28 & 58 \\ 0 & 0 & 0 & 0 \end{bmatrix} \quad \begin{matrix} \\ \\ \text{Row 3} + \tfrac{1}{2} \text{ Row 2.} \end{matrix}$$

The last matrix is in row-echelon form and has two nonzero rows. Hence rank $\mathbf{A} = 2$, as before. ■

Examples 1–3 illustrate the following useful theorem (with $p = 3$, $n = 3$, and the rank of the matrix $= 2$).

THEOREM 2 **Linear Independence and Dependence of Vectors**

Consider p vectors that each have n components. Then these vectors are linearly independent if the matrix formed, with these vectors as row vectors, has rank p. However, these vectors are linearly dependent if that matrix has rank less than p.

Further important properties will result from the basic

THEOREM 3 **Rank in Terms of Column Vectors**

The rank r of a matrix **A** *equals the maximum number of linearly independent* ***column*** *vectors of* **A**.

Hence **A** *and its transpose* $\mathbf{A}^\mathsf{T}$ *have the same rank.*

PROOF In this proof we write simply "rows" and "columns" for row and column vectors. Let **A** be an $m \times n$ matrix of rank $\mathbf{A} = r$. Then by definition of rank, **A** has r linearly independent rows which we denote by $\mathbf{v}_{(1)}, \cdots, \mathbf{v}_{(r)}$ (regardless of their position in **A**), and all the rows $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(m)}$ of **A** are linear combinations of those, say,

$$\begin{aligned} \mathbf{a}_{(1)} &= c_{11}\mathbf{v}_{(1)} + c_{12}\mathbf{v}_{(2)} + \cdots + c_{1r}\mathbf{v}_{(r)} \\ \mathbf{a}_{(2)} &= c_{21}\mathbf{v}_{(1)} + c_{22}\mathbf{v}_{(2)} + \cdots + c_{2r}\mathbf{v}_{(r)} \\ &\vdots \\ \mathbf{a}_{(m)} &= c_{m1}\mathbf{v}_{(1)} + c_{m2}\mathbf{v}_{(2)} + \cdots + c_{mr}\mathbf{v}_{(r)}. \end{aligned} \tag{3}$$

These are vector equations for rows. To switch to columns, we write (3) in terms of components as n such systems, with $k = 1, \cdots, n$,

$$
(4) \qquad \begin{aligned}
a_{1k} &= c_{11}v_{1k} + c_{12}v_{2k} + \cdots + c_{1r}v_{rk} \\
a_{2k} &= c_{21}v_{1k} + c_{22}v_{2k} + \cdots + c_{2r}v_{rk} \\
&\vdots \\
a_{mk} &= c_{m1}v_{1k} + c_{m2}v_{2k} + \cdots + c_{mr}v_{rk}
\end{aligned}
$$

and collect components in columns. Indeed, we can write (4) as

$$
(5) \qquad \begin{bmatrix} a_{1k} \\ a_{2k} \\ \vdots \\ a_{mk} \end{bmatrix} = v_{1k} \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{m1} \end{bmatrix} + v_{2k} \begin{bmatrix} c_{12} \\ c_{22} \\ \vdots \\ c_{m2} \end{bmatrix} + \cdots + v_{rk} \begin{bmatrix} c_{1r} \\ c_{2r} \\ \vdots \\ c_{mr} \end{bmatrix}
$$

where $k = 1, \cdots, n$. Now the vector on the left is the kth column vector of $\mathbf{A}$. We see that each of these n columns is a linear combination of the same r columns on the right. Hence $\mathbf{A}$ cannot have more linearly independent columns than rows, whose number is rank $\mathbf{A} = r$. Now rows of $\mathbf{A}$ are columns of the transpose $\mathbf{A}^\mathsf{T}$. For $\mathbf{A}^\mathsf{T}$ our conclusion is that $\mathbf{A}^\mathsf{T}$ cannot have more linearly independent columns than rows, so that $\mathbf{A}$ cannot have more linearly independent rows than columns. Together, the number of linearly independent columns of $\mathbf{A}$ must be r, the rank of $\mathbf{A}$. This completes the proof. ■

EXAMPLE 4 **Illustration of Theorem 3**

The matrix in (2) has rank 2. From Example 3 we see that the first two row vectors are linearly independent and by "working backward" we can verify that Row 3 = 6 Row 1 $-\frac{1}{2}$ Row 2. Similarly, the first two columns are linearly independent, and by reducing the last matrix in Example 3 by columns we find that

$$
\text{Column 3} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{2}{3}\,\text{Column 2} \qquad \text{and} \qquad \text{Column 4} = \tfrac{2}{3}\,\text{Column 1} + \tfrac{29}{21}\,\text{Column 2.}
$$
■

Combining Theorems 2 and 3 we obtain

THEOREM 4 **Linear Dependence of Vectors**

Consider p vectors each having n components. If $n < p$, then these vectors are linearly dependent.

PROOF The matrix $\mathbf{A}$ with those p vectors as row vectors has p rows and $n < p$ columns; hence by Theorem 3 it has rank $\mathbf{A} \leqq n < p$, which implies linear dependence by Theorem 2. ■

Vector Space

The following related concepts are of general interest in linear algebra. In the present context they provide a clarification of essential properties of matrices and their role in connection with linear systems.

Consider a nonempty set V of vectors where each vector has the same number of components. If, for any two vectors **a** and **b** in V, we have that all their linear combinations $\alpha\mathbf{a} + \beta\mathbf{b}$ (α, β any real numbers) are also elements of V, and if, furthermore, **a** and **b** satisfy the laws (3a), (3c), (3d), and (4) in Sec. 7.1, as well as any vectors **a**, **b**, **c** in V satisfy (3b) then V is a vector space. Note that here we wrote laws (3) and (4) of Sec. 7.1 in lowercase letters **a**, **b**, **c**, which is our notation for vectors. More on vector spaces in Sec. 7.9.

The maximum number of linearly independent vectors in V is called the **dimension** of V and is denoted by dim V. Here we assume the dimension to be finite; infinite dimension will be defined in Sec. 7.9.

A linearly independent set in V consisting of a maximum possible number of vectors in V is called a **basis** for V. In other words, any largest possible set of independent vectors in V forms basis for V. That means, if we add one or more vector to that set, the set will be linearly dependent. (See also the beginning of Sec. 7.4 on linear independence and dependence of vectors.) Thus, the number of vectors of a basis for V equals dim V.

The set of all linear combinations of given vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(p)}$ with the same number of components is called the **span** of these vectors. Obviously, a span is a vector space. If in addition, the given vectors $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(p)}$ are linearly independent, then they form a basis for that vector space.

This then leads to another equivalent definition of basis. A set of vectors is a **basis** for a vector space V if (1) the vectors in the set are linearly independent, and if (2) any vector in V can be expressed as a linear combination of the vectors in the set. If (2) holds, we also say that the set of vectors **spans** the vector space V.

By a **subspace** of a vector space V we mean a nonempty subset of V (including V itself) that forms a vector space with respect to the two algebraic operations (addition and scalar multiplication) defined for the vectors of V.

EXAMPLE 5 Vector Space, Dimension, Basis

The span of the three vectors in Example 1 is a vector space of dimension 2. A basis of this vector space consists of any two of those three vectors, for instance, $\mathbf{a}_{(1)}, \mathbf{a}_{(2)}$, or $\mathbf{a}_{(1)}, \mathbf{a}_{(3)}$, etc. ■

We further note the simple

THEOREM 5 Vector Space R^n

The vector space R^n consisting of all vectors with n components (n real numbers) has dimension n.

PROOF A basis of n vectors is $\mathbf{a}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]$, $\mathbf{a}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0]$, $\cdots$, $\mathbf{a}_{(n)} = [0 \quad \cdots \quad 0 \quad 1]$. ■

For a matrix **A**, we call the span of the row vectors the **row space** of **A**. Similarly, the span of the column vectors of **A** is called the **column space** of **A**.

Now, Theorem 3 shows that a matrix **A** has as many linearly independent rows as columns. By the definition of dimension, their number is the dimension of the row space or the column space of **A**. This proves

THEOREM 6 Row Space and Column Space

The row space and the column space of a matrix **A** *have the same dimension, equal to rank* **A**.

Finally, for a given matrix **A** the solution set of the homogeneous system $\mathbf{Ax} = 0$ is a vector space, called the **null space** of **A**, and its dimension is called the **nullity** of **A**. In the next section we motivate and prove the basic relation

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = \text{Number of columns of } \mathbf{A}. \tag{6}$$

PROBLEM SET 7.4

1–10 RANK, ROW SPACE, COLUMN SPACE

Find the rank. Find a basis for the row space. Find a basis for the column space. *Hint.* Row-reduce the matrix and its transpose. (You may omit obvious factors from the vectors of these bases.)

1. $\begin{bmatrix} 4 & -2 & 6 \\ -2 & 1 & -3 \end{bmatrix}$

2. $\begin{bmatrix} a & b \\ b & a \end{bmatrix}$

3. $\begin{bmatrix} 0 & 3 & 5 \\ 3 & 5 & 0 \\ 5 & 0 & 10 \end{bmatrix}$

4. $\begin{bmatrix} 6 & -4 & 0 \\ -4 & 0 & 2 \\ 0 & 2 & 6 \end{bmatrix}$

5. $\begin{bmatrix} 0.2 & -0.1 & 0.4 \\ 0 & 1.1 & -0.3 \\ 0.1 & 0 & -2.1 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & -4 \\ 0 & 4 & 0 \end{bmatrix}$

7. $\begin{bmatrix} 8 & 0 & 4 & 0 \\ 0 & 2 & 0 & 4 \\ 4 & 0 & 2 & 0 \end{bmatrix}$

8. $\begin{bmatrix} 2 & 4 & 8 & 16 \\ 16 & 8 & 4 & 2 \\ 4 & 8 & 16 & 2 \\ 2 & 16 & 8 & 4 \end{bmatrix}$

9. $\begin{bmatrix} 9 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

10. $\begin{bmatrix} 5 & -2 & 1 & 0 \\ -2 & 0 & -4 & 1 \\ 1 & -4 & -11 & 2 \\ 0 & 1 & 2 & 0 \end{bmatrix}$

11. CAS Experiment. Rank. (a) Show experimentally that the $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ with $a_{jk} = j + k - 1$ has rank 2 for any n. (Problem 20 shows $n = 4$.) Try to prove it.

(b) Do the same when $a_{jk} = j + k + c$, where c is any positive integer.

(c) What is rank **A** if $a_{jk} = 2^{j+k-2}$? Try to find other large matrices of low rank independent of n.

12–16 GENERAL PROPERTIES OF RANK

Show the following:

12. rank $\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ = rank **AB**. (Note the order!)

13. rank **A** = rank **B** does *not* imply rank $\mathbf{A}^2$ = rank $\mathbf{B}^2$. (Give a counterexample.)

14. If **A** is not square, either the row vectors or the column vectors of **A** are linearly dependent.

15. If the row vectors of a square matrix are linearly independent, so are the column vectors, and conversely.

16. Give examples showing that the rank of a product of matrices cannot exceed the rank of either factor.

17–25 LINEAR INDEPENDENCE

Are the following sets of vectors linearly independent? Show the details of your work.

17. $[3 \;\; 4 \;\; 0 \;\; 2]$, $[2 \;\; -1 \;\; 3 \;\; 7]$, $[1 \;\; 16 \;\; -12 \;\; -22]$

18. $[1 \;\; \tfrac{1}{2} \;\; \tfrac{1}{3} \;\; \tfrac{1}{4}]$, $[\tfrac{1}{2} \;\; \tfrac{1}{3} \;\; \tfrac{1}{4} \;\; \tfrac{1}{5}]$, $[\tfrac{1}{3} \;\; \tfrac{1}{4} \;\; \tfrac{1}{5} \;\; \tfrac{1}{6}]$, $[\tfrac{1}{4} \;\; \tfrac{1}{5} \;\; \tfrac{1}{6} \;\; \tfrac{1}{7}]$

19. $[0 \;\; 1 \;\; 1]$, $[1 \;\; 1 \;\; 1]$, $[0 \;\; 0 \;\; 1]$

20. $[1 \;\; 2 \;\; 3 \;\; 4]$, $[2 \;\; 3 \;\; 4 \;\; 5]$, $[3 \;\; 4 \;\; 5 \;\; 6]$, $[4 \;\; 5 \;\; 6 \;\; 7]$

21. $[2 \;\; 0 \;\; 0 \;\; 7]$, $[2 \;\; 0 \;\; 0 \;\; 8]$, $[2 \;\; 0 \;\; 0 \;\; 9]$, $[2 \;\; 0 \;\; 1 \;\; 0]$

22. $[0.4 \;\; -0.2 \;\; 0.2]$, $[0 \;\; 0 \;\; 0]$, $[3.0 \;\; -0.6 \;\; 1.5]$

23. $[9 \;\; 8 \;\; 7 \;\; 6 \;\; 5]$, $[9 \;\; 7 \;\; 5 \;\; 3 \;\; 1]$

24. $[4 \;\; -1 \;\; 3]$, $[0 \;\; 8 \;\; 1]$, $[1 \;\; 3 \;\; -5]$, $[2 \;\; 6 \;\; 1]$

25. $[6 \;\; 0 \;\; -1 \;\; 3]$, $[2 \;\; 2 \;\; 5 \;\; 0]$, $[-4 \;\; -4 \;\; -4 \;\; -4]$

26. Linearly independent subset. Beginning with the last of the vectors $[3 \;\; 0 \;\; 1 \;\; 2]$, $[6 \;\; 1 \;\; 0 \;\; 0]$, $[12 \;\; 1 \;\; 2 \;\; 4]$, $[6 \;\; 0 \;\; 2 \;\; 4]$, and $[9 \;\; 0 \;\; 1 \;\; 2]$, omit one after another until you get a linearly independent set.

27–35 VECTOR SPACE

Is the given set of vectors a vector space? Give reasons. If your answer is yes, determine the dimension and find a basis. ($v_1, v_2, \cdots$ denote components.)

27. All vectors in R^3 with $v_1 - v_2 + 2v_3 = 0$

28. All vectors in R^3 with $3v_2 + v_3 = k$

29. All vectors in R^2 with $v_1 \geqq v_2$

30. All vectors in R^n with the first $n - 2$ components zero

31. All vectors in R^5 with positive components

32. All vectors in R^3 with $3v_1 - 2v_2 + v_3 = 0$, $4v_1 + 5v_2 = 0$

33. All vectors in R^3 with $3v_1 - v_3 = 0$, $2v_1 + 3v_2 - 4v_3 = 0$

34. All vectors in R^n with $|v_j| = 1$ for $j = 1, \cdots, n$

35. All vectors in R^4 with $v_1 = 2v_2 = 3v_3 = 4v_4$

7.5 Solutions of Linear Systems: Existence, Uniqueness

Rank, as just defined, gives complete information about existence, uniqueness, and general structure of the solution set of linear systems as follows.

A linear system of equations in n unknowns has a unique solution if the coefficient matrix and the augmented matrix have the same rank n, and infinitely many solutions if that common rank is less than n. The system has no solution if those two matrices have different rank.

To state this precisely and prove it, we shall use the generally important concept of a **submatrix** of **A**. By this we mean any matrix obtained from **A** by omitting some rows or columns (or both). By definition this includes **A** itself (as the matrix obtained by omitting no rows or columns); this is practical.

THEOREM 1

Fundamental Theorem for Linear Systems

(a) Existence. *A linear system of m equations in n unknowns* $x_1, \cdots, x_n$

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m
\end{aligned}
\tag{1}
$$

is **consistent**, *that is, has solutions, if and only if the coefficient matrix* **A** *and the augmented matrix* $\tilde{\mathbf{A}}$ *have the same rank. Here,*

$$
\mathbf{A} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{m1} & \cdots & a_{mn} \end{bmatrix} \quad \textit{and} \quad \tilde{\mathbf{A}} = \left[\begin{array}{ccc|c} a_{11} & \cdots & a_{1n} & b_1 \\ \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdots & \cdot & \cdot \\ a_{m1} & \cdots & a_{mn} & b_m \end{array}\right]
$$

(b) Uniqueness. *The system* (1) *has precisely one solution if and only if this common rank r of* **A** *and* $\tilde{\mathbf{A}}$ *equals n.*

(c) Infinitely many solutions. *If this common rank r is less than n, the system* (1) *has infinitely many solutions. All of these solutions are obtained by determining r suitable unknowns* (*whose submatrix of coefficients must have rank r*) *in terms of the remaining $n - r$ unknowns, to which arbitrary values can be assigned.* (See Example 3 in Sec. 7.3.)

(d) Gauss elimination (Sec. 7.3). *If solutions exist, they can all be obtained by the Gauss elimination.* (This method will automatically reveal whether or not solutions exist; see Sec. 7.3.)

PROOF **(a)** We can write the system (1) in vector form $\mathbf{Ax} = \mathbf{b}$ or in terms of column vectors $\mathbf{c}_{(1)}, \cdots, \mathbf{c}_{(n)}$ of $\mathbf{A}$:

(2) $$\mathbf{c}_{(1)}x_1 + \mathbf{c}_{(2)}x_2 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{b}.$$

$\widetilde{\mathbf{A}}$ is obtained by augmenting $\mathbf{A}$ by a single column $\mathbf{b}$. Hence, by Theorem 3 in Sec. 7.4, rank $\widetilde{\mathbf{A}}$ equals rank $\mathbf{A}$ or rank $\mathbf{A} + 1$. Now if (1) has a solution $\mathbf{x}$, then (2) shows that $\mathbf{b}$ must be a linear combination of those column vectors, so that $\widetilde{\mathbf{A}}$ and $\mathbf{A}$ have the same maximum number of linearly independent column vectors and thus the same rank.

Conversely, if rank $\widetilde{\mathbf{A}} =$ rank $\mathbf{A}$, then $\mathbf{b}$ must be a linear combination of the column vectors of $\mathbf{A}$, say,

(2*) $$\mathbf{b} = \alpha_1\mathbf{c}_{(1)} + \cdots + \alpha_n\mathbf{c}_{(n)}$$

since otherwise rank $\widetilde{\mathbf{A}} =$ rank $\mathbf{A} + 1$. But (2*) means that (1) has a solution, namely, $x_1 = \alpha_1, \cdots, x_n = \alpha_n$, as can be seen by comparing (2*) and (2).

(b) If rank $\mathbf{A} = n$, the n column vectors in (2) are linearly independent by Theorem 3 in Sec. 7.4. We claim that then the representation (2) of $\mathbf{b}$ is unique because otherwise

$$\mathbf{c}_{(1)}x_1 + \cdots + \mathbf{c}_{(n)}x_n = \mathbf{c}_{(1)}\tilde{x}_1 + \cdots + \mathbf{c}_{(n)}\tilde{x}_n.$$

This would imply (take all terms to the left, with a minus sign)

$$(x_1 - \tilde{x}_1)\mathbf{c}_{(1)} + \cdots + (x_n - \tilde{x}_n)\mathbf{c}_{(n)} = \mathbf{0}$$

and $x_1 - \tilde{x}_1 = 0, \cdots, x_n - \tilde{x}_n = 0$ by linear independence. But this means that the scalars $x_1, \cdots, x_n$ in (2) are uniquely determined, that is, the solution of (1) is unique.

(c) If rank $\mathbf{A} =$ rank $\widetilde{\mathbf{A}} = r < n$, then by Theorem 3 in Sec. 7.4 there is a linearly independent set K of r column vectors of $\mathbf{A}$ such that the other $n - r$ column vectors of $\mathbf{A}$ are linear combinations of those vectors. We renumber the columns and unknowns, denoting the renumbered quantities by ˆ, so that $\{\hat{\mathbf{c}}_{(1)}, \cdots, \hat{\mathbf{c}}_{(r)}\}$ is that linearly independent set K. Then (2) becomes

$$\hat{\mathbf{c}}_{(1)}\hat{x}_1 + \cdots + \hat{\mathbf{c}}_{(r)}\hat{x}_r + \hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1} + \cdots + \hat{\mathbf{c}}_{(n)}\hat{x}_n = \mathbf{b},$$

$\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{\mathbf{c}}_{(n)}$ are linear combinations of the vectors of K, and so are the vectors $\hat{x}_{r+1}\hat{\mathbf{c}}_{(r+1)}, \cdots, \hat{x}_n\hat{\mathbf{c}}_{(n)}$. Expressing these vectors in terms of the vectors of K and collecting terms, we can thus write the system in the form

(3) $$\hat{\mathbf{c}}_{(1)}y_1 + \cdots + \hat{\mathbf{c}}_{(r)}y_r = \mathbf{b}$$

with $y_j = \hat{x}_j + \beta_j$, where β_j results from the $n - r$ terms $\hat{\mathbf{c}}_{(r+1)}\hat{x}_{r+1}, \cdots, \hat{\mathbf{c}}_{(n)}\hat{x}_n$; here, $j = 1, \cdots, r$. Since the system has a solution, there are $y_1, \cdots, y_r$ satisfying (3). These scalars are unique since K is linearly independent. Choosing $\hat{x}_{r+1}, \cdots, \hat{x}_n$ fixes the β_j and corresponding $\hat{x}_j = y_j - \beta_j$, where $j = 1, \cdots, r$.

(d) This was discussed in Sec. 7.3 and is restated here as a reminder. ■

The theorem is illustrated in Sec. 7.3. In Example 2 there is a unique solution since rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A} = n = 3$ (as can be seen from the last matrix in the example). In Example 3 we have rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A} = 2 < n = 4$ and can choose x_3 and x_4 arbitrarily. In Example 4 there is no solution because rank $\mathbf{A} = 2 <$ rank $\tilde{\mathbf{A}} = 3$.

Homogeneous Linear System

Recall from Sec. 7.3 that a linear system (1) is called **homogeneous** if all the b_j's are zero, and **nonhomogeneous** if one or several b_j's are not zero. For the homogeneous system we obtain from the Fundamental Theorem the following results.

THEOREM 2

Homogeneous Linear System

A homogeneous linear system

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= 0 \end{aligned} \tag{4}$$

always has the **trivial solution** $x_1 = 0, \cdots, x_n = 0$. *Nontrivial solutions exist if and only if rank* $\mathbf{A} < n$. *If* rank $\mathbf{A} = r < n$, *these solutions, together with* $\mathbf{x} = \mathbf{0}$, *form a vector space* (see Sec. 7.4) *of dimension* $n - r$ called the **solution space** of (4).

In particular, if $\mathbf{x}_{(1)}$ *and* $\mathbf{x}_{(2)}$ *are solution vectors of* (4), *then* $\mathbf{x} = c_1\mathbf{x}_{(1)} + c_2\mathbf{x}_{(2)}$ *with any scalars* c_1 *and* c_2 *is a solution vector of* (4). (This ***does not hold*** for nonhomogeneous systems. Also, the term *solution space is used for homogeneous* systems only.)

PROOF The first proposition can be seen directly from the system. It agrees with the fact that $\mathbf{b} = \mathbf{0}$ implies that rank $\tilde{\mathbf{A}}$ = rank $\mathbf{A}$, so that a homogeneous system is always ***consistent***. If rank $\mathbf{A} = n$, the trivial solution is the unique solution according to (b) in Theorem 1. If rank $\mathbf{A} < n$, there are nontrivial solutions according to (c) in Theorem 1. The solutions form a vector space because if $\mathbf{x}_{(1)}$ and $\mathbf{x}_{(2)}$ are any of them, then $\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, $\mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$, and this implies $\mathbf{A}(\mathbf{x}_{(1)} + \mathbf{x}_{(2)}) = \mathbf{A}\mathbf{x}_{(1)} + \mathbf{A}\mathbf{x}_{(2)} = \mathbf{0}$ as well as $\mathbf{A}(c\mathbf{x}_{(1)}) = c\mathbf{A}\mathbf{x}_{(1)} = \mathbf{0}$, where c is arbitrary. If rank $\mathbf{A} = r < n$, Theorem 1 (c) implies that we can choose $n - r$ suitable unknowns, call them $x_{r+1}, \cdots, x_n$, in an arbitrary fashion, and every solution is obtained in this way. Hence a basis for the solution space, briefly called a **basis of solutions** of (4), is $\mathbf{y}_{(1)}, \cdots, \mathbf{y}_{(n-r)}$, where the basis vector $\mathbf{y}_{(j)}$ is obtained by choosing $x_{r+j} = 1$ and the other $x_{r+1}, \cdots, x_n$ zero; the corresponding first r components of this solution vector are then determined. Thus the solution space of (4) has dimension $n - r$. This proves Theorem 2. ■

The solution space of (4) is also called the **null space** of $\mathbf{A}$ because $\mathbf{Ax} = \mathbf{0}$ for every $\mathbf{x}$ in the solution space of (4). Its dimension is called the **nullity** of $\mathbf{A}$. Hence Theorem 2 states that

$$\text{rank } \mathbf{A} + \text{nullity } \mathbf{A} = n \qquad \textbf{(5)}$$

where n is the number of unknowns (number of columns of $\mathbf{A}$).

Furthermore, by the definition of rank we have rank $\mathbf{A} \leqq m$ in (4). Hence if $m < n$, then rank $\mathbf{A} < n$. By Theorem 2 this gives the practically important

THEOREM 3

Homogeneous Linear System with Fewer Equations Than Unknowns

A homogeneous linear system with fewer equations than unknowns always has nontrivial solutions.

Nonhomogeneous Linear Systems

The characterization of all solutions of the linear system (1) is now quite simple, as follows.

THEOREM 4

Nonhomogeneous Linear System

If a nonhomogeneous linear system (1) *is consistent, then all of its solutions are obtained as*

$$\mathbf{x} = \mathbf{x}_0 + \mathbf{x}_h \qquad (6)$$

where $\mathbf{x}_0$ *is any* (fixed) *solution of* (1) *and* $\mathbf{x}_h$ *runs through all the solutions of the corresponding homogeneous system* (4).

PROOF The difference $\mathbf{x}_h = \mathbf{x} - \mathbf{x}_0$ of any two solutions of (1) is a solution of (4) because $\mathbf{Ax}_h = \mathbf{A}(\mathbf{x} - \mathbf{x}_0) = \mathbf{Ax} - \mathbf{Ax}_0 = \mathbf{b} - \mathbf{b} = \mathbf{0}$. Since $\mathbf{x}$ is any solution of (1), we get all the solutions of (1) if in (6) we take any solution $\mathbf{x}_0$ of (1) and let $\mathbf{x}_h$ vary throughout the solution space of (4). ■

This covers a main part of our discussion of characterizing the solutions of systems of linear equations. Our next main topic is determinants and their role in linear equations.

7.6 For Reference: Second- and Third-Order Determinants

We created this section as a quick general reference section on second- and third-order determinants. It is completely independent of the theory in Sec. 7.7 and suffices as a reference for many of our examples and problems. Since this section is for reference, ***go on to the next section, consulting this material only when needed***.

A **determinant of second order** is denoted and defined by

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \qquad (1)$$

So here we have ***bars*** (whereas a matrix has ***brackets***).

Cramer's rule for solving linear systems of two equations in two unknowns

$$(2)\qquad \begin{aligned} &\text{(a)}\quad a_{11}x_1 + a_{12}x_2 = b_1 \\ &\text{(b)}\quad a_{21}x_1 + a_{22}x_2 = b_2 \end{aligned}$$

is

$$(3)\qquad \begin{aligned} x_1 &= \frac{\begin{vmatrix} b_1 & a_{12} \\ b_2 & a_{22} \end{vmatrix}}{D} = \frac{b_1a_{22} - a_{12}b_2}{D}, \\ x_2 &= \frac{\begin{vmatrix} a_{11} & b_1 \\ a_{21} & b_2 \end{vmatrix}}{D} = \frac{a_{11}b_2 - b_1a_{21}}{D} \end{aligned}$$

with D as in (1), provided

$$D \neq 0.$$

The value $D = 0$ appears for homogeneous systems with nontrivial solutions.

PROOF We prove (3). To eliminate x_2 multiply (2a) by a_{22} and (2b) by $-a_{12}$ and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_1 = b_1a_{22} - a_{12}b_2.$$

Similarly, to eliminate x_1 multiply (2a) by $-a_{21}$ and (2b) by a_{11} and add,

$$(a_{11}a_{22} - a_{12}a_{21})x_2 = a_{11}b_2 - b_1a_{21}.$$

Assuming that $D = a_{11}a_{22} - a_{12}a_{21} \neq 0$, dividing, and writing the right sides of these two equations as determinants, we obtain (3). ∎

EXAMPLE 1 Cramer's Rule for Two Equations

If $\begin{aligned} 4x_1 + 3x_2 &= 12 \\ 2x_1 + 5x_2 &= -8 \end{aligned}$ then $x_1 = \dfrac{\begin{vmatrix} 12 & 3 \\ -8 & 5 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{84}{14} = 6,\quad x_2 = \dfrac{\begin{vmatrix} 4 & 12 \\ 2 & -8 \end{vmatrix}}{\begin{vmatrix} 4 & 3 \\ 2 & 5 \end{vmatrix}} = \dfrac{-56}{14} = -4.$ ∎

Third-Order Determinants

A **determinant of third order** can be defined by

$$(4)\qquad D = \begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} - a_{21}\begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix} + a_{31}\begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix}.$$

Note the following. The signs on the right are $+ - +$. Each of the three terms on the right is an entry in the first column of D times its **minor**, that is, the second-order determinant obtained from D by deleting the row and column of that entry; thus, for a_{11} delete the first row and first column, and so on.

If we write out the minors in (4), we obtain

$$D = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{21}a_{13}a_{32} - a_{21}a_{12}a_{33} + a_{31}a_{12}a_{23} - a_{31}a_{13}a_{22}. \tag{4*}$$

Cramer's Rule for Linear Systems of Three Equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 &= b_1 \\ a_{21}x_1 + a_{22}x_2 + a_{23}x_3 &= b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 &= b_3 \end{aligned} \tag{5}$$

is

$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \quad x_3 = \frac{D_3}{D} \qquad (D \neq 0) \tag{6}$$

with the *determinant D of the system* given by (4) and

$$D_1 = \begin{vmatrix} b_1 & a_{12} & a_{13} \\ b_2 & a_{22} & a_{23} \\ b_3 & a_{32} & a_{33} \end{vmatrix}, \quad D_2 = \begin{vmatrix} a_{11} & b_1 & a_{13} \\ a_{21} & b_2 & a_{23} \\ a_{31} & b_3 & a_{33} \end{vmatrix}, \quad D_3 = \begin{vmatrix} a_{11} & a_{12} & b_1 \\ a_{21} & a_{22} & b_2 \\ a_{31} & a_{32} & b_3 \end{vmatrix}.$$

Note that D_1, D_2, D_3 are obtained by replacing Columns 1, 2, 3, respectively, by the column of the right sides of (5).

Cramer's rule (6) can be derived by eliminations similar to those for (3), but it also follows from the general case (Theorem 4) in the next section.

7.7 Determinants. Cramer's Rule

Determinants were originally introduced for solving linear systems. Although ***impractical in computations***, they have important engineering applications in eigenvalue problems (Sec. 8.1), differential equations, vector algebra (Sec. 9.3), and in other areas. They can be introduced in several equivalent ways. Our definition is particularly for dealing with linear systems.

A **determinant of order** n is a scalar associated with an $n \times n$ (hence ***square!***) matrix $\mathbf{A} = [a_{jk}]$, and is denoted by

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix}. \tag{1}$$

For $n = 1$, this determinant is defined by

$$D = a_{11}. \tag{2}$$

For $n \geqq 2$ by

$$D = a_{j1}C_{j1} + a_{j2}C_{j2} + \cdots + a_{jn}C_{jn} \qquad (j = 1, 2, \cdots, \text{or } n) \tag{3a}$$

or

$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk} \quad (k = 1, 2, \cdots, \text{or } n). \tag{3b}$$

Here,

$$C_{jk} = (-1)^{j+k} M_{jk}$$

and M_{jk} is a determinant of order $n - 1$, namely, the determinant of the submatrix of **A** obtained from **A** by omitting the row and column of the entry a_{jk}, that is, the jth row and the kth column.

In this way, D is defined in terms of n determinants of order $n - 1$, each of which is, in turn, defined in terms of $n - 1$ determinants of order $n - 2$, and so on—until we finally arrive at second-order determinants, in which those submatrices consist of single entries whose determinant is defined to be the entry itself.

From the definition it follows that *we may **expand** D by any row or column,* that is, choose in (3) the entries in any row or column, similarly when expanding the C_{jk}'s in (3), and so on.

This definition is unambiguous, that is, it yields the same value for D no matter which columns or rows we choose in expanding. A proof is given in App. 4.

Terms used in connection with determinants are taken from matrices. In D we have n^2 **entries** a_{jk}, also n **rows** and n **columns**, and a **main diagonal** on which $a_{11}, a_{22}, \cdots, a_{nn}$ stand. Two terms are new:

M_{jk} is called the **minor** *of* a_{jk} *in* D, and C_{jk} the **cofactor** *of* a_{jk} *in* D.

For later use we note that (3) may also be written in terms of minors

$$D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (j = 1, 2, \cdots, \text{or } n) \tag{4a}$$

$$D = \sum_{j=1}^{n} (-1)^{j+k} a_{jk} M_{jk} \qquad (k = 1, 2, \cdots, \text{or } n). \tag{4b}$$

EXAMPLE 1 Minors and Cofactors of a Third-Order Determinant

In (4) of the previous section the minors and cofactors of the entries in the first column can be seen directly. For the entries in the second row the minors are

$$M_{21} = \begin{vmatrix} a_{12} & a_{13} \\ a_{32} & a_{33} \end{vmatrix}, \qquad M_{22} = \begin{vmatrix} a_{11} & a_{13} \\ a_{31} & a_{33} \end{vmatrix}, \qquad M_{23} = \begin{vmatrix} a_{11} & a_{12} \\ a_{31} & a_{32} \end{vmatrix}$$

and the cofactors are $C_{21} = -M_{21}$, $C_{22} = +M_{22}$, and $C_{23} = -M_{23}$. Similarly for the third row—write these down yourself. And verify that the signs in C_{jk} form a **checkerboard pattern**

$$\begin{matrix} + & - & + \\ - & + & - \\ + & - & + \end{matrix}$$

■

EXAMPLE 2 **Expansions of a Third-Order Determinant**

$$D = \begin{vmatrix} 1 & 3 & 0 \\ 2 & 6 & 4 \\ -1 & 0 & 2 \end{vmatrix} = 1\begin{vmatrix} 6 & 4 \\ 0 & 2 \end{vmatrix} - 3\begin{vmatrix} 2 & 4 \\ -1 & 2 \end{vmatrix} + 0\begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix}$$

$$= 1(12 - 0) - 3(4 + 4) + 0(0 + 6) = -12.$$

This is the expansion by the first row. The expansion by the third column is

$$D = 0\begin{vmatrix} 2 & 6 \\ -1 & 0 \end{vmatrix} - 4\begin{vmatrix} 1 & 3 \\ -1 & 0 \end{vmatrix} + 2\begin{vmatrix} 1 & 3 \\ 2 & 6 \end{vmatrix} = 0 - 12 + 0 = -12.$$

Verify that the other four expansions also give the value -12. ∎

EXAMPLE 3 **Determinant of a Triangular Matrix**

$$\begin{vmatrix} -3 & 0 & 0 \\ 6 & 4 & 0 \\ -1 & 2 & 5 \end{vmatrix} = -3\begin{vmatrix} 4 & 0 \\ 2 & 5 \end{vmatrix} = -3 \cdot 4 \cdot 5 = -60.$$

Inspired by this, can you formulate a little theorem on determinants of triangular matrices? Of diagonal matrices? ∎

General Properties of Determinants

There is an attractive way of finding determinants (1) that consists of applying elementary row operations to (1). By doing so we obtain an "upper triangular" determinant (see Sec. 7.1, for definition with "matrix" replaced by "determinant") whose value is then very easy to compute, being just the product of its diagonal entries. This approach is ***similar*** (***but not the same*!**) to what we did to matrices in Sec. 7.3. *In particular, be aware that interchanging two rows in a determinant introduces a multiplicative factor of* -1 *to the value of the determinant*! Details are as follows.

THEOREM 1 **Behavior of an *n*th-Order Determinant under Elementary Row Operations**

(a) *Interchange of two rows multiplies the value of the determinant by* -1.

(b) *Addition of a multiple of a row to another row does not alter the value of the determinant.*

(c) *Multiplication of a row by a nonzero constant c multiplies the value of the determinant by c.* (This holds also when $c = 0$, but no longer gives an elementary row operation.)

PROOF **(a)** By induction. The statement holds for $n = 2$ because

$$\begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc, \qquad \text{but} \qquad \begin{vmatrix} c & d \\ a & b \end{vmatrix} = bc - ad.$$

We now make the induction hypothesis that (a) holds for determinants of order $n - 1 \geqq 2$ and show that it then holds for determinants of order n. Let D be of order n. Let E be obtained from D by the interchange of two rows. Expand D and E by a row that is ***not*** one of those interchanged, call it the jth row. Then by (4a),

$$D = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} M_{jk}, \qquad E = \sum_{k=1}^{n} (-1)^{j+k} a_{jk} N_{jk} \tag{5}$$

where N_{jk} is obtained from the minor M_{jk} of a_{jk} in D by the interchange of those two rows which have been interchanged in D (and which N_{jk} must both contain because we expand by another row!). Now these minors are of order $n - 1$. Hence the induction hypothesis applies and gives $N_{jk} = -M_{jk}$. Thus $E = -D$ by (5).

(b) Add c times Row i to Row j. Let $\tilde{D}$ be the new determinant. Its entries in Row j are $a_{jk} + ca_{ik}$. If we expand $\tilde{D}$ by this Row j, we see that we can write it as $\tilde{D} = D_1 + cD_2$, where $D_1 = D$ has in Row j the a_{jk}, whereas D_2 has in that Row j the a_{jk} from the addition. Hence D_2 has a_{jk} in both Row i and Row j. Interchanging these two rows gives D_2 back, but on the other hand it gives $-D_2$ by (a). Together $D_2 = -D_2 = 0$, so that $\tilde{D} = D_1 = D$.

(c) Expand the determinant by the row that has been multiplied.

CAUTION! $\det(c\mathbf{A}) = c^n \det \mathbf{A}$ (not $c \det \mathbf{A}$). Explain why. ■

EXAMPLE 4 Evaluation of Determinants by Reduction to Triangular Form

Because of Theorem 1 we may evaluate determinants by reduction to triangular form, as in the Gauss elimination for a matrix. For instance (with the blue explanations always referring to the ***preceding determinant***)

$$D = \begin{vmatrix} 2 & 0 & -4 & 6 \\ 4 & 5 & 1 & 0 \\ 0 & 2 & 6 & -1 \\ -3 & 8 & 9 & 1 \end{vmatrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 2 & 6 & -1 \\ 0 & 8 & 3 & 10 \end{vmatrix} \quad \begin{matrix} \\ \text{Row 2} - 2 \text{ Row 1} \\ \\ \text{Row 4} + 1.5 \text{ Row 1} \end{matrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -11.4 & 29.2 \end{vmatrix} \quad \begin{matrix} \\ \\ \text{Row 3} - 0.4 \text{ Row 2} \\ \text{Row 4} - 1.6 \text{ Row 2} \end{matrix}$$

$$= \begin{vmatrix} 2 & 0 & -4 & 6 \\ 0 & 5 & 9 & -12 \\ 0 & 0 & 2.4 & 3.8 \\ 0 & 0 & -0 & 47.25 \end{vmatrix} \quad \begin{matrix} \\ \\ \\ \text{Row 4} + 4.75 \text{ Row 3} \end{matrix}$$

$$= 2 \cdot 5 \cdot 2.4 \cdot 47.25 = 1134.$$ ■

THEOREM 2

Further Properties of nth-Order Determinants

(a)–(c) *in Theorem* 1 *hold also for columns.*

(d) ***Transposition*** *leaves the value of a determinant unaltered.*

(e) ***A zero row or column*** *renders the value of a determinant zero.*

(f) ***Proportional rows or columns*** *render the value of a determinant zero. In particular, a determinant with two identical rows or columns has the value zero.*

PROOF **(a)–(e)** follow directly from the fact that a determinant can be expanded by any row column. In **(d)**, transposition is defined as for matrices, that is, the jth row becomes the jth column of the transpose.

(f) If Row $j = c$ times Row i, then $D = cD_1$, where D_1 has Row $j =$ Row i. Hence an interchange of these rows reproduces D_1, but it also gives $-D_1$ by Theorem 1(a). Hence $D_1 = 0$ and $D = cD_1 = 0$. Similarly for columns. ∎

It is quite remarkable that the important concept of the rank of a matrix $\mathbf{A}$, which is the maximum number of linearly independent row or column vectors of $\mathbf{A}$ (see Sec. 7.4), can be related to determinants. Here we may assume that rank $\mathbf{A} > 0$ because the only matrices with rank 0 are the zero matrices (see Sec. 7.4).

THEOREM 3

Rank in Terms of Determinants

Consider an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$:

(1) *$\mathbf{A}$ has rank $r \geqq 1$ if and only if $\mathbf{A}$ has an $r \times r$ submatrix with a nonzero determinant.*

(2) *The determinant of any square submatrix with more than r rows, contained in $\mathbf{A}$ (if such a matrix exists!) has a value equal to zero.*

Furthermore, if $m = n$, we have:

(3) *An $n \times n$ square matrix $\mathbf{A}$ has rank n if and only if*

$$\det \mathbf{A} \neq 0.$$

PROOF The key idea is that elementary row operations (Sec. 7.3) alter neither rank (by Theorem 1 in Sec. 7.4) nor the property of a determinant being nonzero (by Theorem 1 in this section). The echelon form $\hat{\mathbf{A}}$ of $\mathbf{A}$ (see Sec. 7.3) has r nonzero row vectors (which are the first r row vectors) if and only if rank $\mathbf{A} = r$. Without loss of generality, we can assume that $r \geqq 1$. Let $\hat{\mathbf{R}}$ be the $r \times r$ submatrix in the left upper corner of $\hat{\mathbf{A}}$ (so that the entries of $\hat{\mathbf{R}}$ are in both the first r rows and r columns of $\hat{\mathbf{A}}$). Now $\hat{\mathbf{R}}$ is triangular, with all diagonal entries r_{jj} nonzero. Thus, $\det \hat{\mathbf{R}} = r_{11} \cdots r_{rr} \neq 0$. Also $\det \mathbf{R} \neq 0$ for the corresponding $r \times r$ submatrix $\mathbf{R}$ of $\mathbf{A}$ because $\hat{\mathbf{R}}$ results from $\mathbf{R}$ by elementary row operations. This proves part (1).

Similarly, $\det \mathbf{S} = 0$ for any square submatrix $\mathbf{S}$ of $r + 1$ or more rows perhaps contained in $\mathbf{A}$ because the corresponding submatrix $\hat{\mathbf{S}}$ of $\hat{\mathbf{A}}$ must contain a row of zeros (otherwise we would have rank $\mathbf{A} \geqq r + 1$), so that $\det \hat{\mathbf{S}} = 0$ by Theorem 2. This proves part (2). Furthermore, we have proven the theorem for an $m \times n$ matrix.

For an $n \times n$ square matrix **A** we proceed as follows. To prove (3), we apply part (1) (already proven!). This gives us that rank $\mathbf{A} = n \geqq 1$ if and only if **A** contains an $n \times n$ submatrix with nonzero determinant. But the only such submatrix contained in our square matrix **A**, is **A** itself, hence $\det \mathbf{A} \neq 0$. This proves part (3). ■

Cramer's Rule

Theorem 3 opens the way to the classical solution formula for linear systems known as **Cramer's rule**,[2] which gives solutions as quotients of determinants. ***Cramer's rule is not practical in computations*** for which the methods in Secs. 7.3 and 20.1–20.3 are suitable. However, Cramer's rule is of ***theoretical interest*** in differential equations (Secs. 2.10 and 3.3) and in other theoretical work that has engineering applications.

THEOREM 4

Cramer's Theorem (Solution of Linear Systems by Determinants)

(a) *If a linear system of n equations in the same number of unknowns* $x_1, \cdots, x_n$

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2 \\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot & \cdot\ \cdot \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n \end{aligned} \tag{6}$$

has a nonzero coefficient determinant $D = \det \mathbf{A}$, *the system has precisely one solution. This solution is given by the formulas*

$$x_1 = \frac{D_1}{D}, \quad x_2 = \frac{D_2}{D}, \cdots, \quad x_n = \frac{D_n}{D} \qquad \textbf{(Cramer's rule)} \tag{7}$$

where D_k *is the determinant obtained from D by replacing in D the kth column by the column with the entries* $b_1, \cdots, b_n$.

(b) *Hence if the system* (6) *is* ***homogeneous*** *and* $D \neq 0$, *it has only the trivial solution* $x_1 = 0, x_2 = 0, \cdots, x_n = 0$. *If* $D = 0$, *the homogeneous system also has nontrivial solutions.*

PROOF The augmented matrix $\tilde{\mathbf{A}}$ of the system (6) is of size $n \times (n + 1)$. Hence its rank can be at most n. Now if

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \cdot & \cdots & \cdot \\ \cdot & \cdots & \cdot \\ a_{n1} & \cdots & a_{nn} \end{vmatrix} \neq 0, \tag{8}$$

[2] GABRIEL CRAMER (1704–1752), Swiss mathematician.

then rank $\mathbf{A} = n$ by Theorem 3. Thus rank $\tilde{\mathbf{A}} = \text{rank}\,\mathbf{A}$. Hence, by the Fundamental Theorem in Sec. 7.5, the system (6) has a unique solution.

Let us now prove (7). Expanding D by its kth column, we obtain

$$D = a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}, \tag{9}$$

where C_{ik} is the cofactor of entry a_{ik} in D. If we replace the entries in the kth column of D by any other numbers, we obtain a new determinant, say, $\hat{D}$. Clearly, its expansion by the kth column will be of the form (9), with $a_{1k}, \cdots, a_{nk}$ replaced by those new numbers and the cofactors C_{ik} as before. In particular, if we choose as new numbers the entries $a_{1l}, \cdots, a_{nl}$ of the lth column of D (where $l \neq k$), we have a new determinant $\hat{D}$ which has the column $[a_{1l} \quad \cdots \quad a_{nl}]^\mathsf{T}$ twice, once as its lth column, and once as its kth because of the replacement. Hence $\hat{D} = 0$ by Theorem 2(f). If we now expand $\hat{D}$ by the column that has been replaced (the kth column), we thus obtain

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk} = 0 \qquad (l \neq k). \tag{10}$$

We now multiply the first equation in (6) by C_{1k} on both sides, the second by $C_{2k}, \cdots,$ the last by C_{nk}, and add the resulting equations. This gives

$$\begin{aligned} C_{1k}(a_{11}x_1 + \cdots + a_{1n}x_n) + \cdots + C_{nk}(a_{n1}x_1 + \cdots + a_{nn}x_n) \\ = b_1C_{1k} + \cdots + b_nC_{nk}. \end{aligned} \tag{11}$$

Collecting terms with the same x_j, we can write the left side as

$$x_1(a_{11}C_{1k} + a_{21}C_{2k} + \cdots + a_{n1}C_{nk}) + \cdots + x_n(a_{1n}C_{1k} + a_{2n}C_{2k} + \cdots + a_{nn}C_{nk}).$$

From this we see that x_k is multiplied by

$$a_{1k}C_{1k} + a_{2k}C_{2k} + \cdots + a_{nk}C_{nk}.$$

Equation (9) shows that this equals D. Similarly, x_1 is multiplied by

$$a_{1l}C_{1k} + a_{2l}C_{2k} + \cdots + a_{nl}C_{nk}.$$

Equation (10) shows that this is zero when $l \neq k$. Accordingly, the left side of (11) equals simply x_kD, so that (11) becomes

$$x_kD = b_1C_{1k} + b_2C_{2k} + \cdots + b_nC_{nk}.$$

Now the right side of this is D_k as defined in the theorem, expanded by its kth column, so that division by D gives (7). This proves Cramer's rule.

If (6) is homogeneous and $D \neq 0$, then each D_k has a column of zeros, so that $D_k = 0$ by Theorem 2(e), and (7) gives the trivial solution.

Finally, if (6) is homogeneous and $D = 0$, then rank $\mathbf{A} < n$ by Theorem 3, so that nontrivial solutions exist by Theorem 2 in Sec. 7.5. ■

EXAMPLE 5 Illustration of Cramer's Rule (Theorem 4)

For $n = 2$, see Example 1 of Sec. 7.6. Also, at the end of that section, we give Cramer's rule for a general linear system of three equations. ■

Finally, an important application for Cramer's rule dealing with inverse matrices will be given in the next section.

PROBLEM SET 7.7

1–6 GENERAL PROBLEMS

1. **General Properties of Determinants.** Illustrate each statement in Theorems 1 and 2 with an example of your choice.

2. **Second-Order Determinant.** Expand a general second-order determinant in four possible ways and show that the results agree.

3. **Third-Order Determinant.** Do the task indicated in Theorem 2. Also evaluate D by reduction to triangular form.

4. **Expansion Numerically Impractical.** Show that the computation of an nth-order determinant by expansion involves $n!$ multiplications, which if a multiplication takes 10^{-9} sec would take these times:

n	10	15	20	25
Time	0.004 sec	22 min	77 years	$0.5 \cdot 10^9$ years

5. **Multiplication by Scalar.** Show that $\det(k\mathbf{A}) = k^n \det \mathbf{A}$ (not $k \det \mathbf{A}$). Give an example.

6. **Minors, cofactors.** Complete the list in Example 1.

7–15 EVALUATION OF DETERMINANTS

Showing the details, evaluate:

7. $\begin{vmatrix} \cos\alpha & \sin\alpha \\ \sin\beta & \cos\beta \end{vmatrix}$

8. $\begin{vmatrix} 0.4 & 4.9 \\ 1.5 & -1.3 \end{vmatrix}$

9. $\begin{vmatrix} \cos n\theta & \sin n\theta \\ -\sin n\theta & \cos n\theta \end{vmatrix}$

10. $\begin{vmatrix} \cosh t & \sinh t \\ \sinh t & \cosh t \end{vmatrix}$

11. $\begin{vmatrix} 4 & -1 & 8 \\ 0 & 2 & 3 \\ 0 & 0 & 5 \end{vmatrix}$

12. $\begin{vmatrix} a & b & c \\ c & a & b \\ b & c & a \end{vmatrix}$

13. $\begin{vmatrix} 0 & 4 & -1 & 5 \\ -4 & 0 & 3 & -2 \\ 1 & -3 & 0 & 1 \\ -5 & 2 & -1 & 0 \end{vmatrix}$

14. $\begin{vmatrix} 4 & 7 & 0 & 0 \\ 2 & 8 & 0 & 0 \\ 0 & 0 & 1 & 5 \\ 0 & 0 & -2 & 2 \end{vmatrix}$

15. $\begin{vmatrix} 1 & 2 & 0 & 0 \\ 2 & 4 & 2 & 0 \\ 0 & 2 & 9 & 2 \\ 0 & 0 & 2 & 16 \end{vmatrix}$

16. **CAS EXPERIMENT. Determinant of Zeros and Ones.** Find the value of the determinant of the $n \times n$ matrix A_n with main diagonal entries all 0 and all others 1. Try to find a formula for this. Try to prove it by induction. Interpret A_3 and A_4 as *incidence matrices* (as in Problem Set 7.1 but without the minuses) of a triangle and a tetrahedron, respectively; similarly for an *n-simplex*, having n vertices and $n(n-1)/2$ edges (and spanning R^{n-1}, $n = 5, 6, \cdots$).

17–19 RANK BY DETERMINANTS

Find the rank by Theorem 3 (which is not very practical) and check by row reduction. Show details.

17. $\begin{bmatrix} 4 & 9 \\ -8 & -6 \\ 16 & 12 \end{bmatrix}$

18. $\begin{bmatrix} 0 & 4 & -6 \\ 4 & 0 & 10 \\ -6 & 10 & 0 \end{bmatrix}$

19. $\begin{bmatrix} 1 & 5 & 2 & 2 \\ 1 & 3 & 2 & 6 \\ 4 & 0 & 8 & 48 \end{bmatrix}$

20. **TEAM PROJECT. Geometric Applications: Curves and Surfaces Through Given Points.** The idea is to get an equation from the vanishing of the determinant of a homogeneous linear system as the condition for a nontrivial solution in Cramer's theorem. We explain the trick for obtaining such a system for the case of a line L through two given points P_1: (x_1, y_1) and P_2: (x_2, y_2). The unknown line is $ax + by = -c$, say. We write it as $ax + by + c \cdot 1 = 0$. To get a nontrivial solution a, b, c, the determinant of the "coefficients" $x, y, 1$ must be zero. The system is

$$\begin{aligned} ax + by + c \cdot 1 &= 0 \quad (\text{Line } L) \\ ax_1 + by_1 + c \cdot 1 &= 0 \quad (P_1 \text{ on } L) \\ ax_2 + by_2 + c \cdot 1 &= 0 \quad (P_2 \text{ on } L). \end{aligned} \tag{12}$$

(a) Line through two points. Derive from $D = 0$ in (12) the familiar formula

$$\frac{x - x_1}{x_1 - x_2} = \frac{y - y_1}{y_1 - y_2}.$$

(b) Plane. Find the analog of (12) for a plane through three given points. Apply it when the points are (1, 1, 1), (3, 2, 6), (5, 0, 5).

(c) Circle. Find a similar formula for a circle in the plane through three given points. Find and sketch the circle through (2, 6), (6, 4), (7, 1).

(d) Sphere. Find the analog of the formula in (c) for a sphere through four given points. Find the sphere through (0, 0, 5), (4, 0, 1), (0, 4, 1), (0, 0, −3) by this formula or by inspection.

(e) General conic section. Find a formula for a general conic section (the vanishing of a determinant of 6th order). Try it out for a quadratic parabola and for a more general conic section of your own choice.

21–25 CRAMER'S RULE

Solve by Cramer's rule. Check by Gauss elimination and back substitution. Show details.

21. $3x - 5y = 15.5$
$6x + 16y = 5.0$

22. $2x - 4y = -24$
$5x + 2y = 0$

23. $3y - 4z = 16$
$2x - 5y + 7z = -27$
$-x - 9z = 9$

24. $3x - 2y + z = 13$
$-2x + y + 4z = 11$
$x + 4y - 5z = -31$

25. $-4w + x + y = -10$
$w - 4x + z = 1$
$w - 4y + z = -7$
$x + y - 4z = 10$

7.8 Inverse of a Matrix. Gauss–Jordan Elimination

***In this section we consider* square *matrices exclusively*.**

The **inverse** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is denoted by $\mathbf{A}^{-1}$ and is an $n \times n$ matrix such that

$$\mathbf{AA}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I} \tag{1}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix (see Sec. 7.2).

If $\mathbf{A}$ has an inverse, then $\mathbf{A}$ is called a **nonsingular matrix**. If $\mathbf{A}$ has no inverse, then $\mathbf{A}$ is called a **singular matrix**.

If $\mathbf{A}$ *has an inverse, the inverse is unique.*

Indeed, if both $\mathbf{B}$ and $\mathbf{C}$ are inverses of $\mathbf{A}$, then $\mathbf{AB} = \mathbf{I}$ and $\mathbf{CA} = \mathbf{I}$, so that we obtain the uniqueness from

$$\mathbf{B} = \mathbf{IB} = (\mathbf{CA})\mathbf{B} = \mathbf{C}(\mathbf{AB}) = \mathbf{CI} = \mathbf{C}.$$

We prove next that $\mathbf{A}$ has an inverse (is nonsingular) if and only if it has maximum possible rank n. The proof will also show that $\mathbf{Ax} = \mathbf{b}$ implies $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ provided $\mathbf{A}^{-1}$ exists, and will thus give a motivation for the inverse as well as a relation to linear systems. (But this will ***not*** give a good method of solving $\mathbf{Ax} = \mathbf{b}$ ***numerically*** because the Gauss elimination in Sec. 7.3 requires fewer computations.)

THEOREM 1

Existence of the Inverse

The inverse $\mathbf{A}^{-1}$ *of an* $n \times n$ *matrix* $\mathbf{A}$ *exists if and only if* rank $\mathbf{A} = n$, *thus* (by Theorem 3, Sec. 7.7) *if and only if* det $\mathbf{A} \neq 0$. *Hence* $\mathbf{A}$ *is nonsingular if* rank $\mathbf{A} = n$, *and is singular if* rank $\mathbf{A} < n$.

PROOF Let $\mathbf{A}$ be a given $n \times n$ matrix and consider the linear system

$$\mathbf{Ax} = \mathbf{b}. \tag{2}$$

If the inverse $\mathbf{A}^{-1}$ exists, then multiplication from the left on both sides and use of (1) gives

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{b}.$$

This shows that (2) has a solution $\mathbf{x}$, which is unique because, for another solution $\mathbf{u}$, we have $\mathbf{Au} = \mathbf{b}$, so that $\mathbf{u} = \mathbf{A}^{-1}\mathbf{b} = \mathbf{x}$. Hence $\mathbf{A}$ must have rank n by the Fundamental Theorem in Sec. 7.5.

Conversely, let rank $\mathbf{A} = n$. Then by the same theorem, the system (2) has a unique solution $\mathbf{x}$ for any $\mathbf{b}$. Now the back substitution following the Gauss elimination (Sec. 7.3) shows that the components x_j of $\mathbf{x}$ are linear combinations of those of $\mathbf{b}$. Hence we can write

$$\mathbf{x} = \mathbf{Bb} \tag{3}$$

with $\mathbf{B}$ to be determined. Substitution into (2) gives

$$\mathbf{Ax} = \mathbf{A}(\mathbf{Bb}) = (\mathbf{AB})\mathbf{b} = \mathbf{Cb} = \mathbf{b} \qquad (\mathbf{C} = \mathbf{AB})$$

for any $\mathbf{b}$. Hence $\mathbf{C} = \mathbf{AB} = \mathbf{I}$, the unit matrix. Similarly, if we substitute (2) into (3) we get

$$\mathbf{x} = \mathbf{Bb} = \mathbf{B}(\mathbf{Ax}) = (\mathbf{BA})\mathbf{x}$$

for any $\mathbf{x}$ (and $\mathbf{b} = \mathbf{Ax}$). Hence $\mathbf{BA} = \mathbf{I}$. Together, $\mathbf{B} = \mathbf{A}^{-1}$ exists. ■

Determination of the Inverse by the Gauss–Jordan Method

To actually determine the inverse $\mathbf{A}^{-1}$ of a nonsingular $n \times n$ matrix $\mathbf{A}$, we can use a variant of the Gauss elimination (Sec. 7.3), called the **Gauss–Jordan elimination**.[3] The idea of the method is as follows.

Using $\mathbf{A}$, we form n linear systems

$$\mathbf{Ax}_{(1)} = \mathbf{e}_{(1)}, \quad \cdots, \quad \mathbf{Ax}_{(n)} = \mathbf{e}_{(n)}$$

where the vectors $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ are the columns of the $n \times n$ unit matrix $\mathbf{I}$; thus, $\mathbf{e}_{(1)} = [1 \quad 0 \quad \cdots \quad 0]^\mathsf{T}$, $\mathbf{e}_{(2)} = [0 \quad 1 \quad 0 \quad \cdots \quad 0]^\mathsf{T}$, etc. These are n vector equations in the unknown vectors $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. We combine them into a single matrix equation

---

[3]WILHELM JORDAN (1842–1899), German geodesist and mathematician. He did important geodesic work in Africa, where he surveyed oases. [See Althoen, S.C. and R. McLaughlin, Gauss–Jordan reduction: A brief history. *American Mathematical Monthly*, Vol. **94**, No. 2 (1987), pp. 130–142.]

We do ***not recommend*** it as a method for solving systems of linear equations, since the number of operations in addition to those of the Gauss elimination is larger than that for back substitution, which the Gauss–Jordan elimination avoids. See also Sec. 20.1.

$\mathbf{AX} = \mathbf{I}$, with the unknown matrix $\mathbf{X}$ having the columns $\mathbf{x}_{(1)}, \cdots, \mathbf{x}_{(n)}$. Correspondingly, we combine the n augmented matrices $[\mathbf{A} \quad \mathbf{e}_{(1)}], \cdots, [\mathbf{A} \quad \mathbf{e}_{(n)}]$ into one wide $n \times 2n$ "augmented matrix" $\widetilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. Now multiplication of $\mathbf{AX} = \mathbf{I}$ by $\mathbf{A}^{-1}$ from the left gives $\mathbf{X} = \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}$. Hence, to solve $\mathbf{AX} = \mathbf{I}$ for $\mathbf{X}$, we can apply the Gauss elimination to $\widetilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{I}]$. This gives a matrix of the form $[\mathbf{U} \quad \mathbf{H}]$ with upper triangular $\mathbf{U}$ because the Gauss elimination triangularizes systems. The Gauss–Jordan method reduces $\mathbf{U}$ by further elementary row operations to diagonal form, in fact to the unit matrix $\mathbf{I}$. This is done by eliminating the entries of $\mathbf{U}$ above the main diagonal and making the diagonal entries all 1 by multiplication (see Example 1). Of course, the method operates on the entire matrix $[\mathbf{U} \quad \mathbf{H}]$, transforming $\mathbf{H}$ into some matrix $\mathbf{K}$, hence the entire $[\mathbf{U} \quad \mathbf{H}]$ to $[\mathbf{I} \quad \mathbf{K}]$. This is the "augmented matrix" of $\mathbf{IX} = \mathbf{K}$. Now $\mathbf{IX} = \mathbf{X} = \mathbf{A}^{-1}$, as shown before. By comparison, $\mathbf{K} = \mathbf{A}^{-1}$, so that we can read $\mathbf{A}^{-1}$ directly from $[\mathbf{I} \quad \mathbf{K}]$.

The following example illustrates the practical details of the method.

EXAMPLE 1 Finding the Inverse of a Matrix by Gauss–Jordan Elimination

Determine the inverse $\mathbf{A}^{-1}$ of

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}.$$

Solution. We apply the Gauss elimination (Sec. 7.3) to the following $n \times 2n = 3 \times 6$ matrix, where BLUE always refers to the previous matrix.

$$[\mathbf{A} \quad \mathbf{I}] = \left[\begin{array}{rrr|rrr} -1 & 1 & 2 & 1 & 0 & 0 \\ 3 & -1 & 1 & 0 & 1 & 0 \\ -1 & 3 & 4 & 0 & 0 & 1 \end{array}\right]$$

$$\left[\begin{array}{rrr|rrr} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 2 & 2 & -1 & 0 & 1 \end{array}\right] \quad \begin{array}{l} \\ \text{Row 2} + 3 \text{ Row 1} \\ \text{Row 3} - \text{Row 1} \end{array}$$

$$\left[\begin{array}{rrr|rrr} -1 & 1 & 2 & 1 & 0 & 0 \\ 0 & 2 & 7 & 3 & 1 & 0 \\ 0 & 0 & -5 & -4 & -1 & 1 \end{array}\right] \quad \begin{array}{l} \\ \\ \text{Row 3} - \text{Row 2} \end{array}$$

This is $[\mathbf{U} \quad \mathbf{H}]$ as produced by the Gauss elimination. Now follow the additional Gauss–Jordan steps, reducing $\mathbf{U}$ to $\mathbf{I}$, that is, to diagonal form with entries 1 on the main diagonal.

$$\left[\begin{array}{rrr|rrr} 1 & -1 & -2 & -1 & 0 & 0 \\ 0 & 1 & 3.5 & 1.5 & 0.5 & 0 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad \begin{array}{l} -\text{Row 1} \\ 0.5 \text{ Row 2} \\ -0.2 \text{ Row 3} \end{array}$$

$$\left[\begin{array}{rrr|rrr} 1 & -1 & 0 & 0.6 & 0.4 & -0.4 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad \begin{array}{l} \text{Row 1} + 2 \text{ Row 3} \\ \text{Row 2} - 3.5 \text{ Row 3} \\ \end{array}$$

$$\left[\begin{array}{rrr|rrr} 1 & 0 & 0 & -0.7 & 0.2 & 0.3 \\ 0 & 1 & 0 & -1.3 & -0.2 & 0.7 \\ 0 & 0 & 1 & 0.8 & 0.2 & -0.2 \end{array}\right] \quad \begin{array}{l} \text{Row 1} + \text{Row 2} \\ \\ \end{array}$$

The last three columns constitute $\mathbf{A}^{-1}$. Check:

$$\begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix} \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Hence $\mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$. Similarly, $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. ■

Formulas for Inverses

Since finding the inverse of a matrix is really a problem of solving a system of linear equations, it is not surprising that Cramer's rule (Theorem 4, Sec. 7.7) might come into play. And similarly, as Cramer's rule was useful for theoretical study but not for computation, so too is the explicit formula (4) in the following theorem useful for theoretical considerations but not recommended for actually determining inverse matrices, except for the frequently occurring 2×2 case as given in (4*).

THEOREM 2

Inverse of a Matrix by Determinants

The inverse of a nonsingular $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is given by

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}}[C_{jk}]^{\mathsf{T}} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} C_{11} & C_{21} & \cdots & C_{n1} \\ C_{12} & C_{22} & \cdots & C_{n2} \\ \cdot & \cdot & \cdots & \cdot \\ C_{1n} & C_{2n} & \cdots & C_{nn} \end{bmatrix}, \tag{4}$$

where C_{jk} is the cofactor of a_{jk} in det $\mathbf{A}$ (see Sec. 7.7). (CAUTION! Note well that in $\mathbf{A}^{-1}$, the cofactor C_{jk} occupies the same place as a_{kj} (not a_{jk}) does in $\mathbf{A}$.)

In particular, the inverse of

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \quad \textit{is} \quad \mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} a_{22} & -a_{12} \\ -a_{21} & a_{11} \end{bmatrix}. \tag{4*}$$

PROOF We denote the right side of (4) by $\mathbf{B}$ and show that $\mathbf{BA} = \mathbf{I}$. We first write

$$\mathbf{BA} = \mathbf{G} = [g_{kl}] \tag{5}$$

and then show that $\mathbf{G} = \mathbf{I}$. Now by the definition of matrix multiplication and because of the form of $\mathbf{B}$ in (4), we obtain (CAUTION! C_{sk}, not C_{ks})

$$g_{kl} = \sum_{s=1}^{n} \frac{C_{sk}}{\det \mathbf{A}} a_{sl} = \frac{1}{\det \mathbf{A}}(a_{1l}C_{1k} + \cdots + a_{nl}C_{nk}). \tag{6}$$

Now (9) and (10) in Sec. 7.7 show that the sum $(\cdots)$ on the right is $D = \det \mathbf{A}$ when $l = k$, and is zero when $l \neq k$. Hence

$$g_{kk} = \frac{1}{\det \mathbf{A}} \det \mathbf{A} = 1,$$

$$g_{kl} = 0 \quad (l \neq k).$$

In particular, for $n = 2$ we have in (4), in the first row, $C_{11} = a_{22}$, $C_{21} = -a_{12}$ and, in the second row, $C_{12} = -a_{21}$, $C_{22} = a_{11}$. This gives (4*). ■

The special case $n = 2$ occurs quite frequently in geometric and other applications. You may perhaps want to memorize formula (4*). Example 2 gives an illustration of (4*).

EXAMPLE 2 Inverse of a 2 × 2 Matrix by Determinants

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 2 & 4 \end{bmatrix}, \quad \mathbf{A}^{-1} = \frac{1}{10}\begin{bmatrix} 4 & -1 \\ -2 & 3 \end{bmatrix} = \begin{bmatrix} 0.4 & -0.1 \\ -0.2 & 0.3 \end{bmatrix}$$ ■

EXAMPLE 3 Further Illustration of Theorem 2

Using (4), find the inverse of

$$\mathbf{A} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}.$$

Solution. We obtain $\det \mathbf{A} = -1(-7) - 1 \cdot 13 + 2 \cdot 8 = 10$, and in (4),

$$C_{11} = \begin{vmatrix} -1 & 1 \\ 3 & 4 \end{vmatrix} = -7, \quad C_{21} = -\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} = 2, \quad C_{31} = \begin{vmatrix} 1 & 2 \\ -1 & 1 \end{vmatrix} = 3,$$

$$C_{12} = -\begin{vmatrix} 3 & 1 \\ -1 & 4 \end{vmatrix} = -13, \quad C_{22} = \begin{vmatrix} -1 & 2 \\ -1 & 4 \end{vmatrix} = -2, \quad C_{32} = -\begin{vmatrix} -1 & 2 \\ 3 & 1 \end{vmatrix} = 7,$$

$$C_{13} = \begin{vmatrix} 3 & -1 \\ -1 & 3 \end{vmatrix} = 8, \quad C_{23} = -\begin{vmatrix} -1 & 1 \\ -1 & 3 \end{vmatrix} = 2, \quad C_{33} = \begin{vmatrix} -1 & 1 \\ 3 & -1 \end{vmatrix} = -2,$$

so that by (4), in agreement with Example 1,

$$\mathbf{A}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$ ■

Diagonal matrices $\mathbf{A} = [a_{jk}]$, $a_{jk} = 0$ when $j \neq k$, have an inverse if and only if all $a_{jj} \neq 0$. Then $\mathbf{A}^{-1}$ is diagonal, too, with entries $1/a_{11}, \cdots, 1/a_{nn}$.

PROOF For a diagonal matrix we have in (4)

$$\frac{C_{11}}{D} = \frac{a_{22} \cdots a_{nn}}{a_{11}a_{22} \cdots a_{nn}} = \frac{1}{a_{11}}, \qquad \text{etc.}$$ ■

EXAMPLE 4 **Inverse of a Diagonal Matrix**

Let

$$\mathbf{A} = \begin{bmatrix} -0.5 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Then we obtain the inverse $\mathbf{A}^{-1}$ by inverting each individual diagonal element of $\mathbf{A}$, that is, by taking $1/(-0.5)$, $\frac{1}{4}$, and $\frac{1}{1}$ as the diagonal entries of $\mathbf{A}^{-1}$, that is,

$$\mathbf{A}^{-1} = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 0.25 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$ ■

Products can be inverted by taking the inverse of each factor and multiplying these inverses ***in reverse order***,

(7) $$(\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}.$$

Hence for more than two factors,

(8) $$(\mathbf{AC} \cdots \mathbf{PQ})^{-1} = \mathbf{Q}^{-1}\mathbf{P}^{-1} \cdots \mathbf{C}^{-1}\mathbf{A}^{-1}.$$

PROOF The idea is to start from (1) for $\mathbf{AC}$ instead of $\mathbf{A}$, that is, $\mathbf{AC}(\mathbf{AC})^{-1} = \mathbf{I}$, and multiply it on both sides from the left, first by $\mathbf{A}^{-1}$, which because of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives

$$\begin{aligned} \mathbf{A}^{-1}\mathbf{AC}(\mathbf{AC})^{-1} &= \mathbf{C}(\mathbf{AC})^{-1} \\ &= \mathbf{A}^{-1}\mathbf{I} = \mathbf{A}^{-1}, \end{aligned}$$

and then multiplying this on both sides from the left, this time by $\mathbf{C}^{-1}$ and by using $\mathbf{C}^{-1}\mathbf{C} = \mathbf{I}$,

$$\mathbf{C}^{-1}\mathbf{C}(\mathbf{AC})^{-1} = (\mathbf{AC})^{-1} = \mathbf{C}^{-1}\mathbf{A}^{-1}.$$

This proves (7), and from it, (8) follows by induction. ■

We also note that *the inverse of the inverse is the given matrix*, as you may prove,

(9) $$(\mathbf{A}^{-1})^{-1} = \mathbf{A}.$$

Unusual Properties of Matrix Multiplication. Cancellation Laws

Section 7.2 contains warnings that some properties of matrix multiplication deviate from those for numbers, and we are now able to explain the restricted validity of the so-called **cancellation laws** [2] and [3] below, using rank and inverse, concepts that were not yet

available in Sec. 7.2. The deviations from the usual are of great practical importance and must be carefully observed. They are as follows.

[1] Matrix multiplication is not commutative, that is, in general we have

$$\mathbf{AB} \neq \mathbf{BA}.$$

[2] $\mathbf{AB} = \mathbf{0}$ does not generally imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ (or $\mathbf{BA} = \mathbf{0}$); for example,

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

[3] $\mathbf{AC} = \mathbf{AD}$ does not generally imply $\mathbf{C} = \mathbf{D}$ (even when $\mathbf{A} \neq \mathbf{0}$).

Complete answers to [2] and [3] are contained in the following theorem.

THEOREM 3

Cancellation Laws

Let $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ *be* $n \times n$ *matrices. Then:*

(a) *If* rank $\mathbf{A} = n$ *and* $\mathbf{AB} = \mathbf{AC}$, *then* $\mathbf{B} = \mathbf{C}$.

(b) *If* rank $\mathbf{A} = n$, *then* $\mathbf{AB} = \mathbf{0}$ *implies* $\mathbf{B} = \mathbf{0}$. *Hence if* $\mathbf{AB} = \mathbf{0}$, *but* $\mathbf{A} \neq \mathbf{0}$ *as well as* $\mathbf{B} \neq \mathbf{0}$, *then* rank $\mathbf{A} < n$ *and* rank $\mathbf{B} < n$.

(c) *If* $\mathbf{A}$ *is singular, so are* $\mathbf{BA}$ *and* $\mathbf{AB}$.

PROOF **(a)** The inverse of $\mathbf{A}$ exists by Theorem 1. Multiplication by $\mathbf{A}^{-1}$ from the left gives $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{A}^{-1}\mathbf{AC}$, hence $\mathbf{B} = \mathbf{C}$.

(b) Let rank $\mathbf{A} = n$. Then $\mathbf{A}^{-1}$ exists, and $\mathbf{AB} = \mathbf{0}$ implies $\mathbf{A}^{-1}\mathbf{AB} = \mathbf{B} = \mathbf{0}$. Similarly when rank $\mathbf{B} = n$. This implies the second statement in (b).

($\mathbf{c_1}$) Rank $\mathbf{A} < n$ by Theorem 1. Hence $\mathbf{Ax} = \mathbf{0}$ has nontrivial solutions by Theorem 2 in Sec. 7.5. Multiplication by $\mathbf{B}$ shows that these solutions are also solutions of $\mathbf{BAx} = \mathbf{0}$, so that rank $(\mathbf{BA}) < n$ by Theorem 2 in Sec. 7.5 and $\mathbf{BA}$ is singular by Theorem 1.

($\mathbf{c_2}$) $\mathbf{A}^\mathsf{T}$ is singular by Theorem 2(d) in Sec. 7.7. Hence $\mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ is singular by part (c_1), and is equal to $(\mathbf{AB})^\mathsf{T}$ by (10d) in Sec. 7.2. Hence $\mathbf{AB}$ is singular by Theorem 2(d) in Sec. 7.7. ∎

Determinants of Matrix Products

The determinant of a matrix product $\mathbf{AB}$ or $\mathbf{BA}$ can be written as the product of the determinants of the factors, and it is interesting that $\det \mathbf{AB} = \det \mathbf{BA}$, although $\mathbf{AB} \neq \mathbf{BA}$ in general. The corresponding formula (10) is needed occasionally and can be obtained by Gauss–Jordan elimination (see Example 1) and from the theorem just proved.

THEOREM 4

Determinant of a Product of Matrices

For any $n \times n$ *matrices* $\mathbf{A}$ *and* $\mathbf{B}$,

$$\det(\mathbf{AB}) = \det(\mathbf{BA}) = \det \mathbf{A} \det \mathbf{B}. \tag{10}$$

PROOF If **A** or **B** is singular, so are **AB** and **BA** by Theorem 3(c), and (10) reduces to $0 = 0$ by Theorem 3 in Sec. 7.7.

Now let **A** and **B** be nonsingular. Then we can reduce **A** to a diagonal matrix $\hat{\mathbf{A}} = [a_{jk}]$ by Gauss–Jordan steps. Under these operations, det **A** retains its value, by Theorem 1 in Sec. 7.7, (a) and (b) [not (c)] except perhaps for a sign reversal in row interchanging when pivoting. But the same operations reduce **AB** to $\hat{\mathbf{A}}\mathbf{B}$ with the same effect on det (**AB**). Hence it remains to prove (10) for $\hat{\mathbf{A}}\mathbf{B}$; written out,

$$\hat{\mathbf{A}}\mathbf{B} = \begin{bmatrix} \hat{a}_{11} & 0 & \cdots & 0 \\ 0 & \hat{a}_{22} & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & \hat{a}_{nn} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ & & \vdots & \\ b_{n1} & b_{n2} & \cdots & b_{nn} \end{bmatrix}$$

$$= \begin{bmatrix} \hat{a}_{11}b_{11} & \hat{a}_{11}b_{12} & \cdots & \hat{a}_{11}b_{1n} \\ \hat{a}_{22}b_{21} & \hat{a}_{22}b_{22} & \cdots & \hat{a}_{22}b_{2n} \\ & & \vdots & \\ \hat{a}_{nn}b_{n1} & \hat{a}_{nn}b_{n2} & \cdots & \hat{a}_{nn}b_{nn} \end{bmatrix}.$$

We now take the determinant det ($\hat{\mathbf{A}}\mathbf{B}$). On the right we can take out a factor $\hat{a}_{11}$ from the first row, $\hat{a}_{22}$ from the second, $\cdots$, $\hat{a}_{nn}$ from the nth. But this product $\hat{a}_{11}\hat{a}_{22}\cdots\hat{a}_{nn}$ equals det $\hat{\mathbf{A}}$ because $\hat{\mathbf{A}}$ is diagonal. The remaining determinant is det **B**. This proves (10) for det (**AB**), and the proof for det (**BA**) follows by the same idea. ■

This completes our discussion of linear systems (Secs. 7.3–7.8). Section 7.9 on vector spaces and linear transformations is optional. ***Numeric methods*** are discussed in Secs. 20.1–20.4, which are independent of other sections on numerics.

PROBLEM SET 7.8

1–10 INVERSE

Find the inverse by Gauss–Jordan (or by (4*) if $n = 2$). Check by using (1).

1. $\begin{bmatrix} 1.80 & -2.32 \\ -0.25 & 0.60 \end{bmatrix}$

2. $\begin{bmatrix} \cos 2\theta & \sin 2\theta \\ -\sin 2\theta & \cos 2\theta \end{bmatrix}$

3. $\begin{bmatrix} 0.3 & -0.1 & 0.5 \\ 2 & 6 & 4 \\ 5 & 0 & 9 \end{bmatrix}$

4. $\begin{bmatrix} 0 & 0 & 0.1 \\ 0 & -0.4 & 0 \\ 2.5 & 0 & 0 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 5 & 4 & 1 \end{bmatrix}$

6. $\begin{bmatrix} -4 & 0 & 0 \\ 0 & 8 & 13 \\ 0 & 3 & 5 \end{bmatrix}$

7. $\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

8. $\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$

9. $\begin{bmatrix} 0 & 8 & 0 \\ 0 & 0 & 4 \\ 2 & 0 & 0 \end{bmatrix}$

10. $\begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$

11–18 SOME GENERAL FORMULAS

11. Inverse of the square. Verify $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2$ for **A** in Prob. 1.

12. Prove the formula in Prob. 11.

13. Inverse of the transpose. Verify $(\mathbf{A}^\mathsf{T})^{-1} = (\mathbf{A}^{-1})^\mathsf{T}$ for $\mathbf{A}$ in Prob. 1.

14. Prove the formula in Prob. 13.

15. Inverse of the inverse. Prove that $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$.

16. Rotation. Give an application of the matrix in Prob. 2 that makes the form of the inverse obvious.

17. Triangular matrix. Is the inverse of a triangular matrix always triangular (as in Prob. 5)? Give reason.

18. Row interchange. Same task as in Prob. 16 for the matrix in Prob. 7.

19–20 FORMULA (4)

Formula (4) is occasionally needed in theory. To understand it, apply it and check the result by Gauss–Jordan:

19. In Prob. 3

20. In Prob. 6

7.9 Vector Spaces, Inner Product Spaces, Linear Transformations *Optional*

We have captured the essence of vector spaces in Sec. 7.4. There we dealt with ***special vector spaces*** that arose quite naturally in the context of matrices and linear systems. The elements of these vector spaces, called *vectors*, satisfied rules (3) and (4) of Sec. 7.1 (which were similar to those for numbers). These special vector spaces were generated by *spans,* that is, linear combination of finitely many vectors. Furthermore, each such vector had n real numbers as *components.* Review this material before going on.

We can generalize this idea by taking *all* vectors with n real numbers as components and obtain the very important ***real n-dimensional vector space*** R^n. The vectors are known as "real vectors." Thus, each vector in R^n is an ordered n-tuple of real numbers.

Now we can consider special values for n. For $n = 2$, we obtain R^2, the vector space of all ordered pairs, which correspond to the **vectors in the plane**. For $n = 3$, we obtain R^3, the vector space of all ordered triples, which are the **vectors in 3-space**. These vectors have wide applications in mechanics, geometry, and calculus and are basic to the engineer and physicist.

Similarly, if we take all ordered n-tuples of *complex numbers* as vectors and complex numbers as scalars, we obtain the **complex vector space** C^n, which we shall consider in Sec. 8.5.

Furthermore, there are other sets of practical interest consisting of matrices, functions, transformations, or others for which addition and scalar multiplication can be defined in an almost natural way so that they too form vector spaces.

It is perhaps not too great an intellectual jump to create, from the ***concrete model*** R^n, the ***abstract concept*** of a ***real vector space*** V by taking the basic properties (3) and (4) in Sec. 7.1 as axioms. In this way, the definition of a real vector space arises.

DEFINITION

Real Vector Space

A nonempty set V of elements $\mathbf{a}, \mathbf{b}, \cdots$ is called a **real vector space** (or *real linear space*), and these elements are called **vectors** (regardless of their nature, which will come out from the context or will be left arbitrary) if, in V, there are defined two algebraic operations (called *vector addition* and *scalar multiplication*) as follows.

I. Vector addition associates with every pair of vectors $\mathbf{a}$ and $\mathbf{b}$ of V a unique vector of V, called the *sum* of $\mathbf{a}$ and $\mathbf{b}$ and denoted by $\mathbf{a} + \mathbf{b}$, such that the following axioms are satisfied.

I.1 *Commutativity.* For any two vectors $\mathbf{a}$ and $\mathbf{b}$ of V,

$$\mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a}.$$

I.2 *Associativity.* For any three vectors $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ of V,

$$(\mathbf{a} + \mathbf{b}) + \mathbf{c} = \mathbf{a} + (\mathbf{b} + \mathbf{c}) \qquad (\text{written } \mathbf{a} + \mathbf{b} + \mathbf{c}).$$

I.3 There is a unique vector in V, called the *zero vector* and denoted by $\mathbf{0}$, such that for every $\mathbf{a}$ in V,

$$\mathbf{a} + \mathbf{0} = \mathbf{a}.$$

I.4 For every $\mathbf{a}$ in V there is a unique vector in V that is denoted by $-\mathbf{a}$ and is such that

$$\mathbf{a} + (-\mathbf{a}) = \mathbf{0}.$$

II. Scalar multiplication. The real numbers are called **scalars**. Scalar multiplication associates with every $\mathbf{a}$ in V and every scalar c a unique vector of V, called the *product* of c and $\mathbf{a}$ and denoted by $c\mathbf{a}$ (or $\mathbf{a}c$) such that the following axioms are satisfied.

II.1 *Distributivity.* For every scalar c and vectors $\mathbf{a}$ and $\mathbf{b}$ in V,

$$c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b}.$$

II.2 *Distributivity.* For all scalars c and k and every $\mathbf{a}$ in V,

$$(c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a}.$$

II.3 *Associativity.* For all scalars c and k and every $\mathbf{a}$ in V,

$$c(k\mathbf{a}) = (ck)\mathbf{a} \qquad (\text{written } ck\mathbf{a}).$$

II.4 For every $\mathbf{a}$ in V,

$$1\mathbf{a} = \mathbf{a}.$$

If, in the above definition, we take complex numbers as scalars instead of real numbers, we obtain the axiomatic definition of a **complex vector space**.

Take a look at the axioms in the above definition. Each axiom stands on its own: It is concise, useful, and it expresses a simple property of V. There are as few axioms as possible and together they express ***all*** the desired properties of V. Selecting good axioms is a process of trial and error that often extends over a long period of time. But once agreed upon, axioms become *standard* such as the ones in the definition of a real vector space.

The following concepts related to a vector space are exactly defined as those given in Sec. 7.4. Indeed, a **linear combination** of vectors $a_{(1)}, \cdots, a_{(m)}$ in a vector space V is an expression

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_m \qquad (c_1, \cdots, c_m \text{ any scalars}).$$

These vectors form a **linearly independent set** (briefly, they are called **linearly independent**) if

$$c_1\mathbf{a}_{(1)} + \cdots + c_m\mathbf{a}_{(m)} = \mathbf{0} \tag{1}$$

implies that $c_1 = 0, \cdots, c_m = 0$. Otherwise, if (1) also holds with scalars not all zero, the vectors are called **linearly dependent**.

Note that (1) with $m = 1$ is $c\mathbf{a} = \mathbf{0}$ and shows that a single vector $\mathbf{a}$ is linearly independent if and only if $\mathbf{a} \neq \mathbf{0}$.

V has **dimension *n***, or is ***n*-dimensional**, if it contains a linearly independent set of n vectors, whereas any set of more than n vectors in V is linearly dependent. That set of n linearly independent vectors is called a **basis** for V. Then every vector in V can be written as a linear combination of the basis vectors. Furthermore, for a given basis, this representation is unique (see Prob. 2).

EXAMPLE 1 **Vector Space of Matrices**

The real 2×2 matrices form a four-dimensional real vector space. A basis is

$$\mathbf{B}_{11} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{12} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad \mathbf{B}_{21} = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad \mathbf{B}_{22} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$

because any 2×2 matrix $\mathbf{A} = [a_{jk}]$ has a unique representation $\mathbf{A} = a_{11}\mathbf{B}_{11} + a_{12}\mathbf{B}_{12} + a_{21}\mathbf{B}_{21} + a_{22}\mathbf{B}_{22}$. Similarly, the real $m \times n$ matrices with fixed m and n form an mn-dimensional vector space. What is the dimension of the vector space of all 3×3 skew-symmetric matrices? Can you find a basis? ■

EXAMPLE 2 **Vector Space of Polynomials**

The set of all constant, linear, and quadratic polynomials in x together is a vector space of dimension 3 with basis $\{1, x, x^2\}$ under the usual addition and multiplication by real numbers because these two operations give polynomials not exceeding degree 2. What is the dimension of the vector space of all polynomials of degree not exceeding a given fixed n? Can you find a basis? ■

If a vector space V contains a linearly independent set of n vectors for every n, no matter how large, then V is called **infinite dimensional**, as opposed to a *finite dimensional* (n-dimensional) vector space just defined. An example of an infinite dimensional vector space is the space of all continuous functions on some interval $[a, b]$ of the x-axis, as we mention without proof.

Inner Product Spaces

If $\mathbf{a}$ and $\mathbf{b}$ are vectors in R^n, regarded as column vectors, we can form the product $\mathbf{a}^\mathsf{T}\mathbf{b}$. This is a 1×1 matrix, which we can identify with its single entry, that is, with a number.

This product is called the **inner product** or **dot product** of **a** and **b**. Other notations for it are (**a**, **b**) and **a** • **b**. Thus

$$\mathbf{a}^{\mathsf{T}}\mathbf{b} = (\mathbf{a}, \mathbf{b}) = \mathbf{a} \cdot \mathbf{b} = [a_1 \cdots a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix} = \sum_{i=1}^{n} a_l b_l = a_1 b_1 + \cdots + a_n b_n.$$

We now extend this concept to general real vector spaces by taking basic properties of (**a**, **b**) as axioms for an "abstract inner product" (**a**, **b**) as follows.

DEFINITION

Real Inner Product Space

A real vector space V is called a **real inner product space** (or *real pre-Hilbert*[4] *space*) if it has the following property. With every pair of vectors **a** and **b** in V there is associated a real number, which is denoted by (**a**, **b**) and is called the **inner product** of **a** and **b**, such that the following axioms are satisfied.

I. For all scalars q_1 and q_2 and all vectors **a**, **b**, **c** in V,

$$(q_1\mathbf{a} + q_2\mathbf{b}, \mathbf{c}) = q_1(\mathbf{a}, \mathbf{c}) + q_2(\mathbf{b}, \mathbf{c}) \qquad (\textit{Linearity}).$$

II. For all vectors **a** and **b** in V,

$$(\mathbf{a}, \mathbf{b}) = (\mathbf{b}, \mathbf{a}) \qquad (\textit{Symmetry}).$$

III. For every **a** in V,

$$\left.\begin{aligned} (\mathbf{a}, \mathbf{a}) &\geqq 0, \\ (\mathbf{a}, \mathbf{a}) = 0 \quad &\text{if and only if} \quad \mathbf{a} = \mathbf{0} \end{aligned}\right\} \quad (\textit{Positive-definiteness}).$$

Vectors whose inner product is zero are called **orthogonal**.
The *length* or **norm** of a vector in V is defined by

(2) $$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} \quad (\geqq 0).$$

A vector of norm 1 is called a **unit vector**.

---

[4]DAVID HILBERT (1862–1943), great German mathematician, taught at Königsberg and Göttingen and was the creator of the famous Göttingen mathematical school. He is known for his basic work in algebra, the calculus of variations, integral equations, functional analysis, and mathematical logic. His "Foundations of Geometry" helped the axiomatic method to gain general recognition. His famous 23 problems (presented in 1900 at the International Congress of Mathematicians in Paris) considerably influenced the development of modern mathematics.

If V is finite dimensional, it is actually a so-called *Hilbert space*; see [GenRef7], p. 128, listed in App. 1.

From these axioms and from (2) one can derive the basic inequality

(3) $$|(\mathbf{a}, \mathbf{b})| \leqq \|\mathbf{a}\| \, \|\mathbf{b}\|$$ (*Cauchy–Schwarz*[5] *inequality*).

From this follows

(4) $$\|\mathbf{a} + \mathbf{b}\| \leqq \|\mathbf{a}\| + \|\mathbf{b}\|$$ (*Triangle inequality*).

A simple direct calculation gives

(5) $$\|\mathbf{a} + \mathbf{b}\|^2 + \|\mathbf{a} - \mathbf{b}\|^2 = 2(\|\mathbf{a}\|^2 + \|\mathbf{b}\|^2)$$ (*Parallelogram equality*).

EXAMPLE 3 ***n*-Dimensional Euclidean Space**

R^n with the inner product

(6) $$(\mathbf{a}, \mathbf{b}) = \mathbf{a}^{\mathsf{T}}\mathbf{b} = a_1 b_1 + \cdots + a_n b_n$$

(where both **a** and **b** are *column* vectors) is called the ***n*-dimensional Euclidean space** and is denoted by E^n or again simply by R^n. Axioms I–III hold, as direct calculation shows. Equation (2) gives the "**Euclidean norm**"

(7) $$\|\mathbf{a}\| = \sqrt{(\mathbf{a}, \mathbf{a})} = \sqrt{\mathbf{a}^{\mathsf{T}}\mathbf{a}} = \sqrt{a_1^2 + \cdots + a_n^2}.$$ ■

EXAMPLE 4 **An Inner Product for Functions. Function Space**

The set of all real-valued continuous functions $f(x), g(x), \cdots$ on a given interval $\alpha \leqq x \leqq \beta$ is a real vector space under the usual addition of functions and multiplication by scalars (real numbers). On this "**function space**" we can define an inner product by the integral

(8) $$(f, g) = \int_{\alpha}^{\beta} f(x)\, g(x)\, dx.$$

Axioms I–III can be verified by direct calculation. Equation (2) gives the norm

(9) $$\|f\| = \sqrt{(f, f)} = \sqrt{\int_{\alpha}^{\beta} f(x)^2 \, dx}.$$ ■

Our examples give a first impression of the great generality of the abstract concepts of vector spaces and inner product spaces. Further details belong to more advanced courses (on functional analysis, meaning abstract modern analysis; see [GenRef7] listed in App. 1) and cannot be discussed here. Instead we now take up a related topic where matrices play a central role.

Linear Transformations

Let X and Y be any vector spaces. To each vector **x** in X we assign a unique vector **y** in Y. Then we say that a **mapping** (or **transformation** or **operator**) of X into Y is given. Such a mapping is denoted by a capital letter, say F. The vector **y** in Y assigned to a vector **x** in X is called the **image** of **x** under F and is denoted by $F(x)$ [or $F\mathbf{x}$, without parentheses].

[5]HERMANN AMANDUS SCHWARZ (1843–1921). German mathematician, known by his work in complex analysis (conformal mapping) and differential geometry. For Cauchy see Sec. 2.5.

F is called a **linear mapping** or **linear transformation** if, for all vectors $\mathbf{v}$ and $\mathbf{x}$ in X and scalars c,

$$\begin{aligned} F(\mathbf{v}+\mathbf{x}) &= F(\mathbf{v}) + F(\mathbf{x}) \\ F(c\mathbf{x}) &= cF(\mathbf{x}). \end{aligned} \tag{10}$$

Linear Transformation of Space R^n into Space R^m

From now on we let $X = R^n$ and $Y = R^m$. Then any real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ gives a transformation of R^n into R^m,

$$\mathbf{y} = \mathbf{A}\mathbf{x}. \tag{11}$$

Since $\mathbf{A}(\mathbf{u} + \mathbf{x}) = \mathbf{A}\mathbf{u} + \mathbf{A}\mathbf{x}$ and $\mathbf{A}(c\mathbf{x}) = c\mathbf{A}\mathbf{x}$, this transformation is linear.

We show that, conversely, every linear transformation F of R^n into R^m can be given in terms of an $m \times n$ matrix $\mathbf{A}$, after a basis for R^n and a basis for R^m have been chosen. This can be proved as follows.

Let $\mathbf{e}_{(1)}, \cdots, \mathbf{e}_{(n)}$ be any basis for R^n. Then every $\mathbf{x}$ in R^n has a unique representation

$$\mathbf{x} = x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}.$$

Since F is linear, this representation implies for the image $F(\mathbf{x})$:

$$F(\mathbf{x}) = F(x_1\mathbf{e}_{(1)} + \cdots + x_n\mathbf{e}_{(n)}) = x_1F(\mathbf{e}_{(1)}) + \cdots + x_nF(\mathbf{e}_{(n)}).$$

Hence F is uniquely determined by the images of the vectors of a basis for R^n. We now choose for R^n the "**standard basis**"

$$\mathbf{e}_{(1)} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \mathbf{e}_{(2)} = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \cdots, \quad \mathbf{e}_{(n)} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix} \tag{12}$$

where $\mathbf{e}_{(j)}$ has its jth component equal to 1 and all others 0. We show that we can now determine an $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ such that for every $\mathbf{x}$ in R^n and image $\mathbf{y} = F(\mathbf{x})$ in R^m,

$$\mathbf{y} = F(\mathbf{x}) = \mathbf{A}\mathbf{x}.$$

Indeed, from the image $\mathbf{y}^{(1)} = F(\mathbf{e}_{(1)})$ of $\mathbf{e}_{(1)}$ we get the condition

$$\mathbf{y}^{(1)} = \begin{bmatrix} y_1^{(1)} \\ y_2^{(1)} \\ \vdots \\ y_m^{(1)} \end{bmatrix} = \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2n} \\ \vdots & & \vdots \\ a_{m1} & \cdots & a_{mm} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

from which we can determine the first column of **A**, namely $a_{11} = y_1^{(1)}$, $a_{21} = y_2^{(1)}$, $\cdots$, $a_{m1} = y_m^{(1)}$. Similarly, from the image of $\mathbf{e}_{(2)}$ we get the second column of **A**, and so on. This completes the proof. ■

We say that **A represents** F, or *is a representation of* F, with respect to the bases for R^n and R^m. Quite generally, the purpose of a "**representation**" is the replacement of one object of study by another object whose properties are more readily apparent.

In three-dimensional Euclidean space E^3 the standard basis is usually written $\mathbf{e}_{(1)} = \mathbf{i}$, $\mathbf{e}_{(2)} = \mathbf{j}$, $\mathbf{e}_{(3)} = \mathbf{k}$. Thus,

$$\mathbf{i} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{j} = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \quad \mathbf{k} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}. \tag{13}$$

These are the three unit vectors in the positive directions of the axes of the **Cartesian coordinate system in space**, that is, the usual coordinate system with the same scale of measurement on the three mutually perpendicular coordinate axes.

EXAMPLE 5 Linear Transformations

Interpreted as transformations of Cartesian coordinates in the plane, the matrices

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} -1 & 0 \\ 0 & -1 \end{bmatrix}, \quad \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix}$$

represent a reflection in the line $x_2 = x_1$, a reflection in the x_1-axis, a reflection in the origin, and a stretch (when $a > 1$, or a contraction when $0 < a < 1$) in the x_1-direction, respectively. ■

EXAMPLE 6 Linear Transformations

Our discussion preceding Example 5 is simpler than it may look at first sight. To see this, find **A** representing the linear transformation that maps (x_1, x_2) onto $(2x_1 - 5x_2, 3x_1 + 4x_2)$.

Solution. Obviously, the transformation is

$$\begin{aligned} y_1 &= 2x_1 - 5x_2 \\ y_2 &= 3x_1 + 4x_2. \end{aligned}$$

From this we can directly see that the matrix is

$$\mathbf{A} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix}. \qquad \text{Check:} \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 & -5 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 2x_1 - 5x_2 \\ 3x_1 + 4x_2 \end{bmatrix}.$$ ■

If **A** in (11) is square, $n \times n$, then (11) maps R^n into R^n. If this **A** is nonsingular, so that $\mathbf{A}^{-1}$ exists (see Sec. 7.8), then multiplication of (11) by $\mathbf{A}^{-1}$ from the left and use of $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ gives the **inverse transformation**

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{y}. \tag{14}$$

It maps every $\mathbf{y} = \mathbf{y}_0$ onto that **x**, which by (11) is mapped onto $\mathbf{y}_0$. *The inverse of a linear transformation is itself linear,* because it is given by a matrix, as (14) shows.

Composition of Linear Transformations

We want to give you a flavor of how linear transformations in general vector spaces work. You will notice, if you read carefully, that definitions and verifications (Example 7) strictly follow the given rules and you can think your way through the material by going in a slow systematic fashion.

The last operation we want to discuss is composition of linear transformations. Let X, Y, W be general vector spaces. As before, let F be a linear transformation from X to Y. Let G be a linear transformation from W to X. Then we denote, by H, the **composition** of F and G, that is,

$$H = F \circ G = FG = F(G),$$

which means *we take transformation G and then apply transformation F to it* (***in that order!***, *i.e. you go from left to right*).

Now, to give this a more concrete meaning, if we let $\mathbf{w}$ be a vector in W, then $G(\mathbf{w})$ is a vector in X and $F(G(\mathbf{w}))$ is a vector in Y. Thus, H maps W to Y, and we can write

$$H(\mathbf{w}) = (F \circ G)(\mathbf{w}) = (FG)(\mathbf{w}) = F(G(\mathbf{w})), \tag{15}$$

which completes the definition of composition in a general vector space setting. But is composition really linear? To check this we have to verify that H, as defined in (15), obeys the two equations of (10).

EXAMPLE 7 The Composition of Linear Transformations Is Linear

To show that H is indeed linear we must show that (10) holds. We have, for two vectors $\mathbf{w}_1$, $\mathbf{w}_2$ in W,

$$\begin{aligned} H(\mathbf{w}_1 + \mathbf{w}_2) &= (F \circ G)(\mathbf{w}_1 + \mathbf{w}_2) \\ &= F(G(\mathbf{w}_1 + \mathbf{w}_2)) \\ &= F(G(\mathbf{w}_1) + G(\mathbf{w}_2)) && \text{(by linearity of } G) \\ &= F(G(\mathbf{w}_1)) + F(G(\mathbf{w}_2)) && \text{(by linearity of } F) \\ &= (F \circ G)(\mathbf{w}_1) + (F \circ G)(\mathbf{w}_2) && \text{(by (15))} \\ &= H(\mathbf{w}_1) + H(\mathbf{w}_2) && \text{(by definition of } H). \end{aligned}$$

Similarly, $H(c\mathbf{w}_2) = (F \circ G)(c\mathbf{w}_2) = F(G(c\mathbf{w}_2)) = F(c(G(\mathbf{w}_2))$

$$= cF(G(\mathbf{w}_2)) = c(F \circ G)(\mathbf{w}_2) = cH(\mathbf{w}_2).$$ ■

We defined composition as a linear transformation in a general vector space setting and showed that the composition of linear transformations is indeed linear.

Next we want to relate composition of linear transformations to matrix multiplication.

To do so we let $X = R^n$, $Y = R^m$, and $W = R^p$. This choice of particular vector spaces allows us to represent the linear transformations as matrices and form matrix equations, as was done in (11). Thus F can be represented by a general real $m \times n$ matrix $\mathbf{A} = [a_{jk}]$ and G by an $n \times p$ matrix $\mathbf{B} = [b_{jk}]$. Then we can write for F, with column vectors $\mathbf{x}$ with n entries, and resulting vector $\mathbf{y}$, with m entries

$$\mathbf{y} = \mathbf{A}\mathbf{x} \tag{16}$$

and similarly for G, with column vector $\mathbf{w}$ with p entries,

$$\mathbf{x} = \mathbf{B}\mathbf{w}. \tag{17}$$

Substituting (17) into (16) gives

$$\mathbf{y} = \mathbf{A}\mathbf{x} = \mathbf{A}(\mathbf{B}\mathbf{w}) = (\mathbf{A}\mathbf{B})\mathbf{w} = \mathbf{A}\mathbf{B}\mathbf{w} = \mathbf{C}\mathbf{w} \qquad \text{where } \mathbf{C} = \mathbf{A}\mathbf{B}. \tag{18}$$

This is (15) in a matrix setting, this is, ***we can define the composition of linear transformations in the Euclidean spaces as multiplication by matrices***. Hence, the real $m \times p$ matrix $\mathbf{C}$ represents a linear transformation H which maps R^p to R^n with vector $\mathbf{w}$, a column vector with p entries.

Remarks. Our discussion is similar to the one in Sec. 7.2, where we motivated the "unnatural" matrix multiplication of matrices. Look back and see that our current, more general, discussion is written out there for the case of dimension $m = 2$, $n = 2$, and $p = 2$. (You may want to write out our development by picking small *distinct* dimensions, such as $m = 2$, $n = 3$, and $p = 4$, and writing down the matrices and vectors. This is a trick of the trade of mathematicians in that we like to develop and test theories on smaller examples to see that they work.)

EXAMPLE 8 Linear Transformations. Composition

In Example 5 of Sec. 7.9, let $\mathbf{A}$ be the first matrix and $\mathbf{B}$ be the fourth matrix with $a > 1$. Then, applying $\mathbf{B}$ to a vector $\mathbf{w} = [w_1 \quad w_2]^\mathsf{T}$, stretches the element w_1 by a in the x_1 direction. Next, when we apply $\mathbf{A}$ to the "stretched" vector, we reflect the vector along the line $x_1 = x_2$, resulting in a vector $\mathbf{y} = [w_2 \quad aw_1]^\mathsf{T}$. But this represents, precisely, a geometric description for the composition H of two linear transformations F and G represented by matrices $\mathbf{A}$ and $\mathbf{B}$. We now show that, for this example, our result can be obtained by straightforward matrix multiplication, that is,

$$\mathbf{AB} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} a & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix}$$

and as in (18) calculate

$$\mathbf{ABw} = \begin{bmatrix} 0 & 1 \\ a & 0 \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} w_2 \\ aw_1 \end{bmatrix},$$

which is the same as before. This shows that indeed $\mathbf{AB} = \mathbf{C}$, and we see the composition of linear transformations can be represented by a linear transformation. It also shows that the order of matrix multiplication is important (!). You may want to try applying $\mathbf{A}$ first and then $\mathbf{B}$, resulting in $\mathbf{BA}$. What do you see? Does it make geometric sense? Is it the same result as **AB?** ■

We have learned several abstract concepts such as vector space, inner product space, and linear transformation. *The introduction of such concepts allows engineers and scientists to communicate in a concise and common language.* For example, the concept of a vector space encapsulated a lot of ideas in a very concise manner. For the student, learning such concepts provides a foundation for more advanced studies in engineering.

This concludes Chapter 7. The central theme was *the Gaussian elimination of* Sec. 7.3 from which most of the other concepts and theory flowed. The next chapter again has a central theme, that is, *eigenvalue problems,* an area very rich in applications such as in engineering, modern physics, and other areas.

PROBLEM SET 7.9

1. **Basis.** Find three bases of R^2.
2. **Uniqueness.** Show that the representation $\mathbf{v} = c_1\mathbf{a}_{(1)} + \cdots + c_n\mathbf{a}_{(n)}$ of any given vector in an n-dimensional vector space V in terms of a given basis $\mathbf{a}_{(1)}, \cdots, \mathbf{a}_{(n)}$ for V is unique. ***Hint***. Take two representations and consider the difference.

3–10 VECTOR SPACE

(More problems in Problem Set 9.4.) Is the given set, taken with the usual addition and scalar multiplication, a vector space? Give reason. If your answer is yes, find the dimension and a basis.

3. All vectors in R^3 satisfying $-v_1 + 2v_2 + 3v_3 = 0$, $-4v_1 + v_2 + v_3 = 0$.
4. All skew-symmetric 3×3 matrices.
5. All polynomials in x of degree 4 or less with nonnegative coefficients.
6. All functions $y(x) = a \cos 2x + b \sin 2x$ with arbitrary constants a and b.
7. All functions $y(x) = (ax + b)e^{-x}$ with any constant a and b.
8. All $n \times n$ matrices $\mathbf{A}$ with fixed n and $\det \mathbf{A} = 0$.
9. All 2×2 matrices $[a_{jk}]$ with $a_{11} + a_{22} = 0$.
10. All 3×2 matrices $[a_{jk}]$ with first column any multiple of $[3 \quad 0 \quad -5]^\mathsf{T}$.

11–14 LINEAR TRANSFORMATIONS

Find the inverse transformation. Show the details.

11. $y_1 = 0.5x_1 - 0.5x_2$
    $y_2 = 1.5x_1 - 2.5x_2$
12. $y_1 = 3x_1 + 2x_2$
    $y_2 = 4x_1 + x_2$
13. $y_1 = 5x_1 + 3x_2 - 3x_3$
    $y_2 = 3x_1 + 2x_2 - 2x_3$
    $y_3 = 2x_1 - x_2 + 2x_3$
14. $y_1 = 0.2x_1 - 0.1x_2$
    $y_2 = -0.2x_2 + 0.1x_3$
    $y_3 = 0.1x_1 + 0.1x_3$

15–20 EUCLIDEAN NORM

Find the Euclidean norm of the vectors:

15. $[3 \quad 1 \quad -4]^\mathsf{T}$
16. $[\tfrac{1}{2} \quad \tfrac{1}{3} \quad -\tfrac{1}{2} \quad -\tfrac{1}{3}]^\mathsf{T}$
17. $[1 \quad 0 \quad 0 \quad 1 \quad -1 \quad 0 \quad -1 \quad 1]^\mathsf{T}$
18. $[-4 \quad 8 \quad -1]^\mathsf{T}$
19. $[\tfrac{2}{3} \quad \tfrac{2}{3} \quad \tfrac{1}{3} \quad 0]^\mathsf{T}$
20. $[\tfrac{1}{2} \quad -\tfrac{1}{2} \quad -\tfrac{1}{2} \quad \tfrac{1}{2}]^\mathsf{T}$

21–25 INNER PRODUCT. ORTHOGONALITY

21. **Orthogonality.** For what value(s) of k are the vectors $[2 \quad \tfrac{1}{2} \quad -4 \quad 0]^\mathsf{T}$ and $[5 \quad k \quad 0 \quad \tfrac{1}{4}]^\mathsf{T}$ orthogonal?
22. **Orthogonality.** Find all vectors in R^3 orthogonal to $[2 \quad 0 \quad 1]$. Do they form a vector space?
23. **Triangle inequality.** Verify (4) for the vectors in Probs. 15 and 18.
24. **Cauchy–Schwarz inequality.** Verify (3) for the vectors in Probs. 16 and 19.
25. **Parallelogram equality.** Verify (5) for the first two column vectors of the coefficient matrix in Prob. 13.

CHAPTER 7 REVIEW QUESTIONS AND PROBLEMS

1. What properties of matrix multiplication differ from those of the multiplication of numbers?
2. Let $\mathbf{A}$ be a 100×100 matrix and $\mathbf{B}$ a 100×50 matrix. Are the following expressions defined or not? $\mathbf{A} + \mathbf{B}$, $\mathbf{A}^2$, $\mathbf{B}^2$, $\mathbf{AB}$, $\mathbf{BA}$, $\mathbf{AA}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}\mathbf{A}$, $\mathbf{B}^\mathsf{T}\mathbf{B}$, $\mathbf{BB}^\mathsf{T}$, $\mathbf{B}^\mathsf{T}\mathbf{AB}$. Give reasons.
3. Are there any linear systems without solutions? With one solution? With more than one solution? Give simple examples.
4. Let $\mathbf{C}$ be 10×10 matrix and $\mathbf{a}$ a column vector with 10 components. Are the following expressions defined or not? $\mathbf{Ca}$, $\mathbf{C}^\mathsf{T}\mathbf{a}$, $\mathbf{Ca}^\mathsf{T}$, $\mathbf{aC}$, $\mathbf{a}^\mathsf{T}\mathbf{C}$, $(\mathbf{Ca}^\mathsf{T})^\mathsf{T}$.
5. Motivate the definition of matrix multiplication.
6. Explain the use of matrices in linear transformations.
7. How can you give the rank of a matrix in terms of row vectors? Of column vectors? Of determinants?
8. What is the role of rank in connection with solving linear systems?
9. What is the idea of Gauss elimination and back substitution?
10. What is the inverse of a matrix? When does it exist? How would you determine it?

11–20 MATRIX AND VECTOR CALCULATIONS

Showing the details, calculate the following expressions or give reason why they are not defined, when

$$\mathbf{A} = \begin{bmatrix} 3 & 1 & -3 \\ 1 & 4 & 2 \\ -3 & 2 & 5 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} 0 & 4 & 1 \\ -4 & 0 & -2 \\ -1 & 2 & 0 \end{bmatrix},$$

$$\mathbf{u} = \begin{bmatrix} 2 \\ 0 \\ -5 \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} 7 \\ -3 \\ 3 \end{bmatrix}$$

11. $\mathbf{AB}$, $\mathbf{BA}$

12. $\mathbf{A}^{\mathsf{T}}$, $\mathbf{B}^{\mathsf{T}}$

13. $\mathbf{Au}$, $\mathbf{u}^{\mathsf{T}}\mathbf{A}$

14. $\mathbf{u}^{\mathsf{T}}\mathbf{v}$, $\mathbf{uv}^{\mathsf{T}}$

15. $\mathbf{u}^{\mathsf{T}}\mathbf{Au}$, $\mathbf{v}^{\mathsf{T}}\mathbf{Bv}$

16. $\mathbf{A}^{-1}$, $\mathbf{B}^{-1}$

17. $\det \mathbf{A}$, $\det \mathbf{A}^2$, $(\det \mathbf{A})^2$, $\det \mathbf{B}$

18. $(\mathbf{A}^2)^{-1}$, $(\mathbf{A}^{-1})^2$

19. $\mathbf{AB} - \mathbf{BA}$

20. $(\mathbf{A} + \mathbf{A}^{\mathsf{T}})(\mathbf{B} - \mathbf{B}^{\mathsf{T}})$

21–28 LINEAR SYSTEMS

Showing the details, find all solutions or indicate that no solution exists.

21.
$$\begin{aligned} 4y + z &= 0 \\ 12x - 5y - 3z &= 34 \\ -6x + 4z &= 8 \end{aligned}$$

22.
$$\begin{aligned} 5x - 3y + z &= 7 \\ 2x + 3y - z &= 0 \\ 8x + 9y - 3z &= 2 \end{aligned}$$

23.
$$\begin{aligned} 9x + 3y - 6z &= 60 \\ 2x - 4y + 8z &= 4 \end{aligned}$$

24.
$$\begin{aligned} -6x + 39y - 9z &= -12 \\ 2x - 13y + 3z &= 4 \end{aligned}$$

25.
$$\begin{aligned} 0.3x - 0.7y + 1.3z &= 3.24 \\ 0.9y - 0.8z &= -2.53 \\ 0.7z &= 1.19 \end{aligned}$$

26.
$$\begin{aligned} 2x + 3y - 7z &= 3 \\ -4x - 6y + 14z &= 7 \end{aligned}$$

27.
$$\begin{aligned} x + 2y &= 6 \\ 3x + 5y &= 20 \\ -4x + y &= -42 \end{aligned}$$

28.
$$\begin{aligned} -8x + 2z &= 1 \\ 6y + 4z &= 3 \\ 12x + 2y &= 2 \end{aligned}$$

29–32 RANK

Determine the ranks of the coefficient matrix and the augmented matrix and state how many solutions the linear system will have.

29. In Prob. 23

30. In Prob. 24

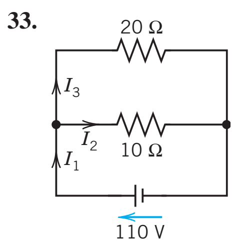
31. In Prob. 27

32. In Prob. 26

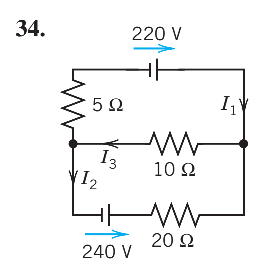
33–35 NETWORKS

Find the currents.

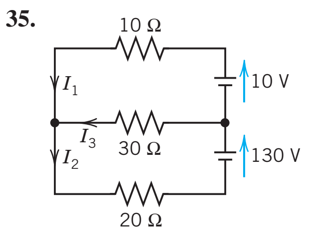
33.



34.



35.



SUMMARY OF CHAPTER 7

Linear Algebra: Matrices, Vectors, Determinants. Linear Systems

An $m \times n$ **matrix** $\mathbf{A} = [a_{jk}]$ is a rectangular array of numbers or functions ("entries," "elements") arranged in m horizontal **rows** and n vertical **columns**. If $m = n$, the matrix is called **square**. A $1 \times n$ matrix is called a **row vector** and an $m \times 1$ matrix a **column vector** (Sec. 7.1).

The **sum** $\mathbf{A} + \mathbf{B}$ of matrices of the same **size** (i.e., both $m \times n$) is obtained by adding corresponding entries. The **product** of $\mathbf{A}$ by a scalar c is obtained by multiplying each a_{jk} by c (Sec. 7.1).

The **product** $\mathbf{C} = \mathbf{AB}$ of an $m \times n$ matrix $\mathbf{A}$ by an $r \times p$ matrix $\mathbf{B} = [b_{jk}]$ is defined only when $r = n$, and is the $m \times p$ matrix $\mathbf{C} = [c_{jk}]$ with entries

$$c_{jk} = a_{j1}b_{1k} + a_{j2}b_{2k} + \cdots + a_{jn}b_{nk} \qquad \text{(row } j \text{ of } \mathbf{A} \text{ times column } k \text{ of } \mathbf{B}\text{)}. \tag{1}$$

This multiplication is motivated by the composition of **linear transformations** (Secs. 7.2, 7.9). It is associative, but is ***not commutative:*** if $\mathbf{AB}$ is defined, $\mathbf{BA}$ may not be defined, but even if $\mathbf{BA}$ is defined, $\mathbf{AB} \neq \mathbf{BA}$ in general. Also $\mathbf{AB} = \mathbf{0}$ may not imply $\mathbf{A} = \mathbf{0}$ or $\mathbf{B} = \mathbf{0}$ or $\mathbf{BA} = \mathbf{0}$ (Secs. 7.2, 7.8). Illustrations:

$$\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}$$

$$[1 \quad 2]\begin{bmatrix} 3 \\ 4 \end{bmatrix} = [11], \qquad \begin{bmatrix} 3 \\ 4 \end{bmatrix}[1 \quad 2] = \begin{bmatrix} 3 & 6 \\ 4 & 8 \end{bmatrix}.$$

The **transpose** $\mathbf{A}^\mathsf{T}$ of a matrix $\mathbf{A} = [a_{jk}]$ is $\mathbf{A}^\mathsf{T} = [a_{kj}]$; rows become columns and conversely (Sec. 7.2). Here, $\mathbf{A}$ need not be square. If it is and $\mathbf{A} = \mathbf{A}^\mathsf{T}$, then $\mathbf{A}$ is called **symmetric;** if $\mathbf{A} = -\mathbf{A}^\mathsf{T}$, it is called **skew-symmetric**. For a product, $(\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T}$ (Sec. 7.2).

A main application of matrices concerns **linear systems of equations**

$$\mathbf{Ax} = \mathbf{b} \qquad \text{(Sec. 7.3)} \tag{2}$$

(m equations in n unknowns $x_1, \cdots, x_n$; $\mathbf{A}$ and $\mathbf{b}$ given). The most important method of solution is the **Gauss elimination** (Sec. 7.3), which reduces the system to "triangular" form by *elementary row operations,* which leave the set of solutions unchanged. (Numeric aspects and variants, such as *Doolittle's* and *Cholesky's methods,* are discussed in Secs. 20.1 and 20.2.)

Cramer's rule (Secs. 7.6, 7.7) represents the unknowns in a system (2) of n equations in n unknowns as quotients of determinants; for numeric work it is impractical. **Determinants** (Sec. 7.7) have decreased in importance, but will retain their place in eigenvalue problems, elementary geometry, etc.

The **inverse** $\mathbf{A}^{-1}$ of a square matrix satisfies $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$. It exists if and only if $\det \mathbf{A} \neq 0$. It can be computed by the *Gauss–Jordan elimination* (Sec. 7.8).

The **rank** r of a matrix $\mathbf{A}$ is the maximum number of linearly independent rows or columns of $\mathbf{A}$ or, equivalently, the number of rows of the largest square submatrix of $\mathbf{A}$ with nonzero determinant (Secs. 7.4, 7.7).

The system (2) has solutions if and only if $\operatorname{rank} \mathbf{A} = \operatorname{rank} [\mathbf{A} \quad \mathbf{b}]$, where $[\mathbf{A} \quad \mathbf{b}]$ is the **augmented matrix** (Fundamental Theorem, Sec. 7.5).

The **homogeneous system**

$$\mathbf{A}\mathbf{x} = \mathbf{0} \tag{3}$$

has solutions $\mathbf{x} \neq \mathbf{0}$ ("nontrivial solutions") if and only if $\operatorname{rank} \mathbf{A} < n$, in the case $m = n$ equivalently if and only if $\det \mathbf{A} = \mathbf{0}$ (Secs. 7.6, 7.7).

Vector spaces, inner product spaces, and linear transformations are discussed in Sec. 7.9. See also Sec. 7.4.

CHAPTER 8

Linear Algebra: Matrix Eigenvalue Problems

A matrix eigenvalue problem considers the vector equation

$$\mathbf{A}\mathbf{x} = \lambda\mathbf{x}. \tag{1}$$

Here **A** is a given square matrix, λ an unknown scalar, and **x** an unknown vector. In a matrix eigenvalue problem, the task is to determine λ's and **x**'s that satisfy (1). Since $\mathbf{x} = \mathbf{0}$ is always a solution for any λ and thus not interesting, we only admit solutions with $\mathbf{x} \neq \mathbf{0}$.

The solutions to (1) are given the following names: The λ's that satisfy (1) are called **eigenvalues of A** and the corresponding nonzero **x**'s that also satisfy (1) are called **eigenvectors of A**.

From this rather innocent looking vector equation flows an amazing amount of relevant theory and an incredible richness of applications. Indeed, eigenvalue problems come up all the time in engineering, physics, geometry, numerics, theoretical mathematics, biology, environmental science, urban planning, economics, psychology, and other areas. Thus, in your career you are likely to encounter eigenvalue problems.

We start with a basic and thorough introduction to eigenvalue problems in Sec. 8.1 and explain (1) with several simple matrices. This is followed by a section devoted entirely to applications ranging from mass–spring systems of physics to population control models of environmental science. We show you these diverse examples to train your skills in modeling and solving eigenvalue problems. Eigenvalue problems for real symmetric, skew-symmetric, and orthogonal matrices are discussed in Sec. 8.3 and their complex counterparts (which are important in modern physics) in Sec. 8.5. In Sec. 8.4 we show how by diagonalizing a matrix, we obtain its eigenvalues.

COMMENT. ***Numerics for eigenvalues* (Secs. 20.6–20.9) *can be studied immediately after this chapter*.**

Prerequisite: Chap. 7.
Sections that may be omitted in a shorter course: 8.4, 8.5.
References and Answers to Problems: App. 1 Part B, App. 2.

The following chart identifies where different types of eigenvalue problems appear in the book.

Topic	Where to find it
Matrix Eigenvalue Problem (algebraic eigenvalue problem)	Chap. 8
Eigenvalue Problems in Numerics	Secs. 20.6–20.9
Eigenvalue Problem for ODEs (Sturm–Liouville problems)	Secs. 11.5, 11.6
Eigenvalue Problems for Systems of ODEs	Chap. 4
Eigenvalue Problems for PDEs	Secs. 12.3–12.11

8.1 The Matrix Eigenvalue Problem. Determining Eigenvalues and Eigenvectors

Consider multiplying nonzero vectors by a given square matrix, such as

$$\begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 5 \\ 1 \end{bmatrix} = \begin{bmatrix} 33 \\ 27 \end{bmatrix}, \qquad \begin{bmatrix} 6 & 3 \\ 4 & 7 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 30 \\ 40 \end{bmatrix}.$$

We want to see what influence the multiplication of the given matrix has on the vectors. In the first case, we get a totally new vector with a different direction and different length when compared to the original vector. This is what usually happens and is of no interest here. In the second case something interesting happens. The multiplication produces a vector $[30 \quad 40]^\mathsf{T} = 10\,[3 \quad 4]^\mathsf{T}$, which means the new vector has the same direction as the original vector. The scale constant, which we denote by λ is 10. *The problem of systematically finding such* λ's *and nonzero vectors for a given square matrix will be the theme of this chapter*. It is called the *matrix eigenvalue* problem or, more commonly, the *eigenvalue* problem.

We formalize our observation. Let $\mathbf{A} = [a_{jk}]$ be a given nonzero square matrix of dimension $n \times n$. Consider the following vector equation:

$$\mathbf{Ax} = \lambda\mathbf{x}. \tag{1}$$

The problem of finding nonzero $\mathbf{x}$'s and λ's that satisfy equation (1) is called an eigenvalue problem.

Remark. So $\mathbf{A}$ is a given square (!) matrix, $\mathbf{x}$ is an unknown vector, and λ is an unknown scalar. Our task is to find λ's and nonzero $\mathbf{x}$'s that satisfy (1). Geometrically, we are looking for vectors, $\mathbf{x}$, for which the multiplication by $\mathbf{A}$ has the same effect as the multiplication by a scalar λ; in other words, $\mathbf{Ax}$ should be proportional to $\mathbf{x}$. Thus, the multiplication has the effect of producing, from the original vector $\mathbf{x}$, a new vector $\lambda\mathbf{x}$ that has the same or opposite (minus sign) direction as the original vector. (This was all demonstrated in our intuitive opening example. Can you see that the second equation in that example satisfies (1) with $\lambda = 10$ and $\mathbf{x} = [3 \quad 4]^\mathsf{T}$, and $\mathbf{A}$ the given 2×2 matrix? Write it out.) Now why do we require $\mathbf{x}$ to be nonzero? The reason is that $\mathbf{x} = \mathbf{0}$ is always a solution of (1) for any value of λ, because $\mathbf{A0} = \mathbf{0}$. This is of no interest.

We introduce more terminology. A value of λ, for which (1) has a solution $\mathbf{x} \neq \mathbf{0}$, is called an **eigenvalue** or *characteristic value* of the matrix $\mathbf{A}$. Another term for λ is a *latent* root. ("Eigen" is German and means "proper" or "characteristic."). The corresponding solutions $\mathbf{x} \neq \mathbf{0}$ of (1) are called the **eigenvectors** or *characteristic vectors* of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all the eigenvalues of $\mathbf{A}$ is called the **spectrum** of $\mathbf{A}$. We shall see that the spectrum consists of at least one eigenvalue and at most of n numerically different eigenvalues. The largest of the absolute values of the eigenvalues of $\mathbf{A}$ is called the *spectral radius* of $\mathbf{A}$, a name to be motivated later.

How to Find Eigenvalues and Eigenvectors

Now, with the new terminology for (1), we can just say that the problem of determining the eigenvalues and eigenvectors of a matrix is called an eigenvalue problem. (However, more precisely, we are considering an algebraic eigenvalue problem, as opposed to an eigenvalue problem involving an ODE or PDE, as considered in Secs. 11.5 and 12.3, or an integral equation.)

Eigenvalues have a very large number of applications in diverse fields such as in engineering, geometry, physics, mathematics, biology, environmental science, economics, psychology, and other areas. You will encounter applications for elastic membranes, Markov processes, population models, and others in this chapter.

Since, from the viewpoint of engineering applications, eigenvalue problems are the most important problems in connection with matrices, the student should carefully follow our discussion.

Example 1 demonstrates how to systematically solve a simple eigenvalue problem.

EXAMPLE 1 Determination of Eigenvalues and Eigenvectors

We illustrate all the steps in terms of the matrix

$$\mathbf{A} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}.$$

Solution. **(a)** ***Eigenvalues.*** These must be determined ***first***. Equation (1) is

$$\mathbf{Ax} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \lambda\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \text{in components,} \qquad \begin{aligned} -5x_1 + 2x_2 &= \lambda x_1 \\ 2x_1 - 2x_2 &= \lambda x_2. \end{aligned}$$

Transferring the terms on the right to the left, we get

$$\text{(2*)} \qquad \begin{aligned} (-5 - \lambda)x_1 + \qquad 2x_2 \qquad &= 0 \\ 2x_1 \qquad + (-2 - \lambda)x_2 &= 0. \end{aligned}$$

This can be written in matrix notation

$$\text{(3*)} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$$

because (1) is $\mathbf{Ax} - \lambda\mathbf{x} = \mathbf{Ax} - \lambda\mathbf{Ix} = (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, which gives (3*). We see that this is a ***homogeneous*** linear system. By Cramer's theorem in Sec. 7.7 it has a nontrivial solution $\mathbf{x} \neq \mathbf{0}$ (an eigenvector of $\mathbf{A}$ we are looking for) if and only if its coefficient determinant is zero, that is,

$$\text{(4*)} \qquad D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -5 - \lambda & 2 \\ 2 & -2 - \lambda \end{vmatrix} = (-5 - \lambda)(-2 - \lambda) - 4 = \lambda^2 + 7\lambda + 6 = 0.$$

We call $D(\lambda)$ the **characteristic determinant** or, if expanded, the **characteristic polynomial**, and $D(\lambda) = 0$ the **characteristic equation** of **A**. The solutions of this quadratic equation are $\lambda_1 = -1$ and $\lambda_2 = -6$. These are the eigenvalues of **A**.

(b_1) ***Eigenvector of* A *corresponding to*** λ_1. This vector is obtained from (2*) with $\lambda = \lambda_1 = -1$, that is,

$$\begin{aligned} -4x_1 + 2x_2 &= 0 \\ 2x_1 - x_2 &= 0. \end{aligned}$$

A solution is $x_2 = 2x_1$, as we see from either of the two equations, so that we need only one of them. This determines an eigenvector corresponding to $\lambda_1 = -1$ up to a scalar multiple. If we choose $x_1 = 1$, we obtain the eigenvector

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \qquad \text{Check:} \qquad \mathbf{A}\mathbf{x}_1 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} -1 \\ -2 \end{bmatrix} = (-1)\mathbf{x}_1 = \lambda_1\mathbf{x}_1.$$

(b_2) ***Eigenvector of* A *corresponding to*** λ_2. For $\lambda = \lambda_2 = -6$, equation (2*) becomes

$$\begin{aligned} x_1 + 2x_2 &= 0 \\ 2x_1 + 4x_2 &= 0. \end{aligned}$$

A solution is $x_2 = -x_1/2$ with arbitrary x_1. If we choose $x_1 = 2$, we get $x_2 = -1$. Thus an eigenvector of **A** corresponding to $\lambda_2 = -6$ is

$$\mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}, \qquad \text{Check:} \qquad \mathbf{A}\mathbf{x}_2 = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix}\begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} -12 \\ 6 \end{bmatrix} = (-6)\mathbf{x}_2 = \lambda_2\mathbf{x}_2.$$

For the matrix in the intuitive opening example at the start of Sec. 8.1, the characteristic equation is $\lambda^2 - 13\lambda + 30 = (\lambda - 10)(\lambda - 3) = 0$. The eigenvalues are $\{10, \ 3\}$. Corresponding eigenvectors are $[3 \quad 4]^T$ and $[-1 \quad 1]^T$, respectively. The reader may want to verify this. ■

This example illustrates the general case as follows. Equation (1) written in components is

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= \lambda x_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= \lambda x_2 \\ &\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n &= \lambda x_n. \end{aligned}$$

Transferring the terms on the right side to the left side, we have

$$(2) \qquad \begin{aligned} (a_{11} - \lambda)x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= 0 \\ a_{21}x_1 + (a_{22} - \lambda)x_2 + \cdots + a_{2n}x_n &= 0 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + (a_{nn} - \lambda)x_n &= 0. \end{aligned}$$

In matrix notation,

$$\textbf{(3)} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

By Cramer's theorem in Sec. 7.7, this homogeneous linear system of equations has a nontrivial solution if and only if the corresponding determinant of the coefficients is zero:

$$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{4}$$

$\mathbf{A} - \lambda\mathbf{I}$ is called the **characteristic matrix** and $D(\lambda)$ the **characteristic determinant** of $\mathbf{A}$. Equation (4) is called the **characteristic equation** of $\mathbf{A}$. By developing $D(\lambda)$ we obtain a polynomial of nth degree in λ. This is called the **characteristic polynomial** of $\mathbf{A}$.

This proves the following important theorem.

THEOREM 1

Eigenvalues

The eigenvalues of a square matrix $\mathbf{A}$ *are the roots of the characteristic equation* (4) *of* $\mathbf{A}$.

Hence an $n \times n$ *matrix has at least one eigenvalue and at most* n *numerically different eigenvalues.*

For larger n, the actual computation of eigenvalues will, in general, require the use of Newton's method (Sec. 19.2) or another numeric approximation method in Secs. 20.7–20.9.

The eigenvalues must be determined first. Once these are known, corresponding eigen*vectors* are obtained from the system (2), for instance, by the Gauss elimination, where λ is the eigenvalue for which an eigenvector is wanted. This is what we did in Example 1 and shall do again in the examples below. (To prevent misunderstandings: *numeric approximation methods*, such as in Sec. 20.8, may determine eigen*vectors first*.)

Eigenvectors have the following properties.

THEOREM 2

Eigenvectors, Eigenspace

If $\mathbf{w}$ *and* $\mathbf{x}$ *are eigenvectors of a matrix* $\mathbf{A}$ *corresponding to* ***the same*** *eigenvalue* λ, *so are* $\mathbf{w} + \mathbf{x}$ (provided $\mathbf{x} \neq -\mathbf{w}$) *and* $k\mathbf{x}$ *for any* $k \neq 0$.

Hence the eigenvectors corresponding to one and the same eigenvalue λ *of* $\mathbf{A}$, *together with* $\mathbf{0}$, *form a vector space* (cf. Sec. 7.4), *called the* **eigenspace** *of* $\mathbf{A}$ *corresponding to that* λ.

PROOF $\mathbf{A}\mathbf{w} = \lambda\mathbf{w}$ and $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$ *imply* $\mathbf{A}(\mathbf{w} + \mathbf{x}) = \mathbf{A}\mathbf{w} + \mathbf{A}\mathbf{x} = \lambda\mathbf{w} + \lambda\mathbf{x} = \lambda(\mathbf{w} + \mathbf{x})$ and $\mathbf{A}(k\mathbf{w}) = k(\mathbf{A}\mathbf{w}) = k(\lambda\mathbf{w}) = \lambda(k\mathbf{w})$; hence $\mathbf{A}(k\mathbf{w} + \ell\mathbf{x}) = \lambda(k\mathbf{w} + \ell\mathbf{x})$. ■

In particular, *an eigenvector* $\mathbf{x}$ *is determined only up to a constant factor*. Hence we can **normalize** $\mathbf{x}$, that is, multiply it by a scalar to get a unit vector (see Sec. 7.9). For instance, $\mathbf{x}_1 = [1 \quad 2]^\mathsf{T}$ in Example 1 has the length $\|\mathbf{x}_1\| = \sqrt{1^2 + 2^2} = \sqrt{5}$; hence $[1/\sqrt{5} \quad 2/\sqrt{5}]^\mathsf{T}$ is a normalized eigenvector (a unit eigenvector).

Examples 2 and 3 will illustrate that an $n \times n$ matrix may have n linearly independent eigenvectors, or it may have fewer than n. In Example 4 we shall see that a *real* matrix may have *complex* eigenvalues and eigenvectors.

EXAMPLE 2 Multiple Eigenvalues

Find the eigenvalues and eigenvectors of

$$\mathbf{A} = \begin{bmatrix} -2 & 2 & -3 \\ 2 & 1 & -6 \\ -1 & -2 & 0 \end{bmatrix}.$$

Solution. For our matrix, the characteristic determinant gives the characteristic equation

$$-\lambda^3 - \lambda^2 + 21\lambda + 45 = 0.$$

The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 5, \lambda_2 = \lambda_3 = -3$. (If you have trouble finding roots, you may want to use a root finding algorithm such as Newton's method (Sec. 19.2). Your CAS or scientific calculator can find roots. However, to really learn and remember this material, you have to do some exercises with paper and pencil.) To find eigenvectors, we apply the Gauss elimination (Sec. 7.3) to the system $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$, first with $\lambda = 5$ and then with $\lambda = -3$. For $\lambda = 5$ the characteristic matrix is

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} - 5\mathbf{I} = \begin{bmatrix} -7 & 2 & -3 \\ 2 & -4 & -6 \\ -1 & -2 & -5 \end{bmatrix}. \quad \text{It row-reduces to} \quad \begin{bmatrix} -7 & 2 & -3 \\ 0 & -\frac{24}{7} & -\frac{48}{7} \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 2. Choosing $x_3 = -1$ we have $x_2 = 2$ from $-\frac{24}{7}x_2 - \frac{48}{7}x_3 = 0$ and then $x_1 = 1$ from $-7x_1 + 2x_2 - 3x_3 = 0$. Hence an eigenvector of $\mathbf{A}$ corresponding to $\lambda = 5$ is $\mathbf{x}_1 = [1 \quad 2 \quad -1]^T$.

For $\lambda = -3$ the characteristic matrix

$$\mathbf{A} - \lambda\mathbf{I} = \mathbf{A} + 3\mathbf{I} = \begin{bmatrix} 1 & 2 & -3 \\ 2 & 4 & -6 \\ -1 & -2 & 3 \end{bmatrix} \quad \text{row-reduces to} \quad \begin{bmatrix} 1 & 2 & -3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

Hence it has rank 1. From $x_1 + 2x_2 - 3x_3 = 0$ we have $x_1 = -2x_2 + 3x_3$. Choosing $x_2 = 1, x_3 = 0$ and $x_2 = 0, x_3 = 1$, we obtain two linearly independent eigenvectors of $\mathbf{A}$ corresponding to $\lambda = -3$ [as they must exist by (5), Sec. 7.5, with rank $= 1$ and $n = 3$],

$$\mathbf{x}_2 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix}$$

and

$$\mathbf{x}_3 = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}.$$

■

The order M_λ of an eigenvalue λ as a root of the characteristic polynomial is called the **algebraic multiplicity** of λ. The number m_λ of linearly independent eigenvectors corresponding to λ is called the **geometric multiplicity** of λ. Thus m_λ is the dimension of the eigenspace corresponding to this λ.

Since the characteristic polynomial has degree n, the sum of all the algebraic multiplicities must equal n. In Example 2 for $\lambda = -3$ we have $m_\lambda = M_\lambda = 2$. In general, $m_\lambda \leqq M_\lambda$, as can be shown. The difference $\Delta_\lambda = M_\lambda - m_\lambda$ is called the **defect** of λ. Thus $\Delta_{-3} = 0$ in Example 2, but positive defects Δ_λ can easily occur:

EXAMPLE 3 Algebraic Multiplicity, Geometric Multiplicity. Positive Defect

The characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \qquad \text{is} \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ 0 & -\lambda \end{vmatrix} = \lambda^2 = 0.$$

Hence $\lambda = 0$ is an eigenvalue of algebraic multiplicity $M_0 = 2$. But its geometric multiplicity is only $m_0 = 1$, since eigenvectors result from $-0x_1 + x_2 = 0$, hence $x_2 = 0$, in the form $[x_1 \quad 0]^\mathsf{T}$. Hence for $\lambda = 0$ the defect is $\Delta_0 = 1$.

Similarly, the characteristic equation of the matrix

$$\mathbf{A} = \begin{bmatrix} 3 & 2 \\ 0 & 3 \end{bmatrix} \qquad \text{is} \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} 3-\lambda & 2 \\ 0 & 3-\lambda \end{vmatrix} = (3-\lambda)^2 = 0.$$

Hence $\lambda = 3$ is an eigenvalue of algebraic multiplicity $M_3 = 2$, but its geometric multiplicity is only $m_3 = 1$, since eigenvectors result from $0x_1 + 2x_2 = 0$ in the form $[x_1 \quad 0]^\mathsf{T}$. ■

EXAMPLE 4 Real Matrices with Complex Eigenvalues and Eigenvectors

Since real polynomials may have complex roots (which then occur in conjugate pairs), a real matrix may have complex eigenvalues and eigenvectors. For instance, the characteristic equation of the skew-symmetric matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \qquad \text{is} \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} -\lambda & 1 \\ -1 & -\lambda \end{vmatrix} = \lambda^2 + 1 = 0.$$

It gives the eigenvalues $\lambda_1 = i\,(= \sqrt{-1})$, $\lambda_2 = -i$. Eigenvectors are obtained from $-ix_1 + x_2 = 0$ and $ix_1 + x_2 = 0$, respectively, and we can choose $x_1 = 1$ to get

$$\begin{bmatrix} 1 \\ i \end{bmatrix} \qquad \text{and} \qquad \begin{bmatrix} 1 \\ -i \end{bmatrix}.$$ ■

In the next section we shall need the following simple theorem.

THEOREM 3 Eigenvalues of the Transpose

The transpose $\mathbf{A}^\mathsf{T}$ of a square matrix $\mathbf{A}$ has the same eigenvalues as $\mathbf{A}$.

PROOF Transposition does not change the value of the characteristic determinant, as follows from Theorem 2d in Sec. 7.7. ■

Having gained a first impression of matrix eigenvalue problems, we shall illustrate their importance with some typical applications in Sec. 8.2.

PROBLEM SET 8.1

1–16 EIGENVALUES, EIGENVECTORS

Find the eigenvalues. Find the corresponding eigenvectors. Use the given λ or factor in Probs. 11 and 15.

1. $\begin{bmatrix} 3.0 & 0 \\ 0 & -0.6 \end{bmatrix}$

2. $\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 5 & -2 \\ 9 & -6 \end{bmatrix}$

4. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

5. $\begin{bmatrix} 0 & 3 \\ -3 & 0 \end{bmatrix}$

6. $\begin{bmatrix} 1 & 2 \\ 0 & 3 \end{bmatrix}$

7. $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$

8. $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

9. $\begin{bmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{bmatrix}$

10. $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

11. $\begin{bmatrix} 6 & 2 & -2 \\ 2 & 5 & 0 \\ -2 & 0 & 7 \end{bmatrix}$, $\lambda = 3$

12. $\begin{bmatrix} 3 & 5 & 3 \\ 0 & 4 & 6 \\ 0 & 0 & 1 \end{bmatrix}$

13. $\begin{bmatrix} 13 & 5 & 2 \\ 2 & 7 & -8 \\ 5 & 4 & 7 \end{bmatrix}$

14. $\begin{bmatrix} 2 & 0 & -1 \\ 0 & \frac{1}{2} & 0 \\ 1 & 0 & 4 \end{bmatrix}$

15. $\begin{bmatrix} -1 & 0 & 12 & 0 \\ 0 & -1 & 0 & 12 \\ 0 & 0 & -1 & -4 \\ 0 & 0 & -4 & -1 \end{bmatrix}$, $(\lambda + 1)^2$

16. $\begin{bmatrix} -3 & 0 & 4 & 2 \\ 0 & 1 & -2 & 4 \\ 2 & 4 & -1 & -2 \\ 0 & 2 & -2 & 3 \end{bmatrix}$

17–20 LINEAR TRANSFORMATIONS AND EIGENVALUES

Find the matrix $\mathbf{A}$ in the linear transformation $\mathbf{y} = \mathbf{Ax}$, where $\mathbf{x} = [x_1 \;\; x_2]^\mathsf{T}$ ($\mathbf{x} = [x_1 \;\; x_2 \;\; x_3]^\mathsf{T}$) are Cartesian coordinates. Find the eigenvalues and eigenvectors and explain their geometric meaning.

17. Counterclockwise rotation through the angle $\pi/2$ about the origin in R^2.

18. Reflection about the x_1-axis in R^2.

19. Orthogonal projection (perpendicular projection) of R^2 onto the x_2-axis.

20. Orthogonal projection of R^3 onto the plane $x_2 = x_1$.

21–25 GENERAL PROBLEMS

21. Nonzero defect. Find further 2×2 and 3×3 matrices with positive defect. See Example 3.

22. Multiple eigenvalues. Find further 2×2 and 3×3 matrices with multiple eigenvalues. See Example 2.

23. Complex eigenvalues. Show that the eigenvalues of a real matrix are real or complex conjugate in pairs.

24. Inverse matrix. Show that $\mathbf{A}^{-1}$ exists if and only if the eigenvalues $\lambda_1, \cdots, \lambda_n$ are all nonzero, and then $\mathbf{A}^{-1}$ has the eigenvalues $1/\lambda_1, \cdots, 1/\lambda_n$.

25. Transpose. Illustrate Theorem 3 with examples of your own.

8.2 Some Applications of Eigenvalue Problems

We have selected some typical examples from the wide range of applications of matrix eigenvalue problems. The last example, that is, Example 4, shows an application involving vibrating springs and ODEs. It falls into the domain of Chapter 4, which covers matrix eigenvalue problems related to ODE's modeling mechanical systems and electrical

networks. Example 4 is included to keep our discussion independent of Chapter 4. (However, the reader not interested in ODEs may want to skip Example 4 without loss of continuity.)

EXAMPLE 1 Stretching of an Elastic Membrane

An elastic membrane in the x_1x_2-plane with boundary circle $x_1^2 + x_2^2 = 1$ (Fig. 160) is stretched so that a point P: (x_1, x_2) goes over into the point Q: (y_1, y_2) given by

$$(1) \qquad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{Ax} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}; \qquad \text{in components,} \qquad \begin{aligned} y_1 &= 5x_1 + 3x_2 \\ y_2 &= 3x_1 + 5x_2. \end{aligned}$$

Find the **principal directions**, that is, the directions of the position vector $\mathbf{x}$ of P for which the direction of the position vector $\mathbf{y}$ of Q is the same or exactly opposite. What shape does the boundary circle take under this deformation?

Solution. We are looking for vectors $\mathbf{x}$ such that $\mathbf{y} = \lambda\mathbf{x}$. Since $\mathbf{y} = \mathbf{Ax}$, this gives $\mathbf{Ax} = \lambda\mathbf{x}$, the equation of an eigenvalue problem. In components, $\mathbf{Ax} = \lambda\mathbf{x}$ is

$$(2) \qquad \begin{aligned} 5x_1 + 3x_2 &= \lambda x_1 \\ 3x_1 + 5x_2 &= \lambda x_2 \end{aligned} \qquad \text{or} \qquad \begin{aligned} (5 - \lambda)x_1 + \quad 3x_2 \quad &= 0 \\ 3x_1 \quad + (5 - \lambda)x_2 &= 0. \end{aligned}$$

The characteristic equation is

$$(3) \qquad \begin{vmatrix} 5 - \lambda & 3 \\ 3 & 5 - \lambda \end{vmatrix} = (5 - \lambda)^2 - 9 = 0.$$

Its solutions are $\lambda_1 = 8$ and $\lambda_2 = 2$. These are the eigenvalues of our problem. For $\lambda = \lambda_1 = 8$, our system (2) becomes

$$\begin{aligned} -3x_1 + 3x_2 &= 0, \\ 3x_1 - 3x_2 &= 0. \end{aligned} \quad \Bigg| \quad \begin{aligned} &\text{Solution } x_2 = x_1, \ x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = x_2 = 1. \end{aligned}$$

For $\lambda_2 = 2$, our system (2) becomes

$$\begin{aligned} 3x_1 + 3x_2 &= 0, \\ 3x_1 + 3x_2 &= 0. \end{aligned} \quad \Bigg| \quad \begin{aligned} &\text{Solution } x_2 = -x_1, \ x_1 \text{ arbitrary,} \\ &\text{for instance, } x_1 = 1, x_2 = -1. \end{aligned}$$

We thus obtain as eigenvectors of $\mathbf{A}$, for instance, $[1 \quad 1]^\mathsf{T}$ corresponding to λ_1 and $[1 \quad -1]^\mathsf{T}$ corresponding to λ_2 (or a nonzero scalar multiple of these). These vectors make 45° and 135° angles with the positive x_1-direction. They give the principal directions, the answer to our problem. The eigenvalues show that in the principal directions the membrane is stretched by factors 8 and 2, respectively; see Fig. 160.

Accordingly, if we choose the principal directions as directions of a new Cartesian u_1u_2-coordinate system, say, with the positive u_1-semi-axis in the first quadrant and the positive u_2-semi-axis in the second quadrant of the x_1x_2-system, and if we set $u_1 = r\cos\phi$, $u_2 = r\sin\phi$, then a boundary point of the unstretched circular membrane has coordinates $\cos\phi$, $\sin\phi$. Hence, after the stretch we have

$$z_1 = 8\cos\phi, \qquad z_2 = 2\sin\phi.$$

Since $\cos^2\phi + \sin^2\phi = 1$, this shows that the deformed boundary is an ellipse (Fig. 160)

$$(4) \qquad \frac{z_1^2}{8^2} + \frac{z_2^2}{2^2} = 1.$$ ■

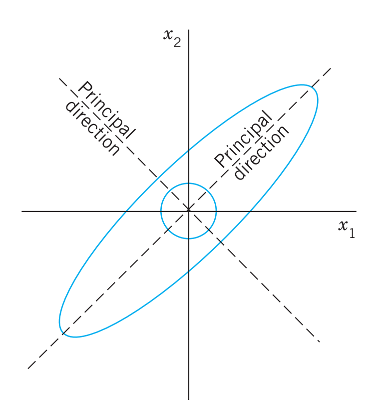


Fig. 160. Undeformed and deformed membrane in Example 1

EXAMPLE 2 Eigenvalue Problems Arising from Markov Processes

Markov processes as considered in Example 13 of Sec. 7.2 lead to eigenvalue problems if we ask for the limit state of the process in which the state vector $\mathbf{x}$ is reproduced under the multiplication by the stochastic matrix $\mathbf{A}$ governing the process, that is, $\mathbf{Ax} = \mathbf{x}$. Hence $\mathbf{A}$ should have the eigenvalue 1, and $\mathbf{x}$ should be a corresponding eigenvector. This is of practical interest because it shows the long-term tendency of the development modeled by the process.

In that example,

$$\mathbf{A} = \begin{bmatrix} 0.7 & 0.1 & 0 \\ 0.2 & 0.9 & 0.2 \\ 0.1 & 0 & 0.8 \end{bmatrix}. \qquad \text{For the transpose,} \qquad \begin{bmatrix} 0.7 & 0.2 & 0.1 \\ 0.1 & 0.9 & 0 \\ 0 & 0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Hence $\mathbf{A}^{\mathsf{T}}$ has the eigenvalue 1, and the same is true for $\mathbf{A}$ by Theorem 3 in Sec. 8.1. An eigenvector $\mathbf{x}$ of $\mathbf{A}$ for $\lambda = 1$ is obtained from

$$\mathbf{A} - \mathbf{I} = \begin{bmatrix} -0.3 & 0.1 & 0 \\ 0.2 & -0.1 & 0.2 \\ 0.1 & 0 & -0.2 \end{bmatrix}, \qquad \text{row-reduced to} \qquad \begin{bmatrix} -\frac{3}{10} & \frac{1}{10} & 0 \\ 0 & -\frac{1}{30} & \frac{1}{5} \\ 0 & 0 & 0 \end{bmatrix}.$$

Taking $x_3 = 1$, we get $x_2 = 6$ from $-x_2/30 + x_3/5 = 0$ and then $x_1 = 2$ from $-3x_1/10 + x_2/10 = 0$. This gives $\mathbf{x} = [2 \quad 6 \quad 1]^{\mathsf{T}}$. It means that in the long run, the ratio Commercial:Industrial:Residential will approach 2:6:1, provided that the probabilities given by $\mathbf{A}$ remain (about) the same. (We switched to ordinary fractions to avoid rounding errors.) ■

EXAMPLE 3 Eigenvalue Problems Arising from Population Models. Leslie Model

The Leslie model describes age-specified population growth, as follows. Let the oldest age attained by the females in some animal population be 9 years. Divide the population into three age classes of 3 years each. Let the "*Leslie matrix*" be

$$\text{(5)} \qquad \mathbf{L} = [l_{jk}] = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix}$$

where l_{1k} is the average number of daughters born to a single female during the time she is in age class k, and $l_{j,j-1}(j = 2, 3)$ is the fraction of females in age class $j - 1$ that will survive and pass into class j. (a) What is the number of females in each class after 3, 6, 9 years if each class initially consists of 400 females? (b) For what initial distribution will the number of females in each class change by the same proportion? What is this rate of change?

Solution. **(a)** Initially, $\mathbf{x}_{(0)}^{\mathsf{T}} = [400 \quad 400 \quad 400]$. After 3 years,

$$\mathbf{x}_{(3)} = \mathbf{L}\mathbf{x}_{(0)} = \begin{bmatrix} 0 & 2.3 & 0.4 \\ 0.6 & 0 & 0 \\ 0 & 0.3 & 0 \end{bmatrix} \begin{bmatrix} 400 \\ 400 \\ 400 \end{bmatrix} = \begin{bmatrix} 1080 \\ 240 \\ 120 \end{bmatrix}.$$

Similarly, after 6 years the number of females in each class is given by $\mathbf{x}_{(6)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(3)})^{\mathsf{T}} = [600 \quad 648 \quad 72]$, and after 9 years we have $\mathbf{x}_{(9)}^{\mathsf{T}} = (\mathbf{L}\mathbf{x}_{(6)})^{\mathsf{T}} = [1519.2 \quad 360 \quad 194.4]$.

(b) Proportional change means that we are looking for a distribution vector $\mathbf{x}$ such that $\mathbf{L}\mathbf{x} = \lambda\mathbf{x}$, where λ is the rate of change (growth if $\lambda > 1$, decrease if $\lambda < 1$). The characteristic equation is (develop the characteristic determinant by the first column)

$$\det(\mathbf{L} - \lambda\mathbf{I}) = -\lambda^3 - 0.6(-2.3\lambda - 0.3 \cdot 0.4) = -\lambda^3 + 1.38\lambda + 0.072 = 0.$$

A positive root is found to be (for instance, by Newton's method, Sec. 19.2) $\lambda = 1.2$. A corresponding eigenvector $\mathbf{x}$ can be determined from the characteristic matrix

$$\mathbf{A} - 1.2\mathbf{I} = \begin{bmatrix} -1.2 & 2.3 & 0.4 \\ 0.6 & -1.2 & 0 \\ 0 & 0.3 & -1.2 \end{bmatrix}, \qquad \text{say,} \qquad \mathbf{x} = \begin{bmatrix} 1 \\ 0.5 \\ 0.125 \end{bmatrix}$$

where $x_3 = 0.125$ is chosen, $x_2 = 0.5$ then follows from $0.3x_2 - 1.2x_3 = 0$, and $x_1 = 1$ from $-1.2x_1 + 2.3x_2 + 0.4x_3 = 0$. To get an initial population of 1200 as before, we multiply $\mathbf{x}$ by $1200/(1 + 0.5 + 0.125) = 738$. *Answer:* Proportional growth of the numbers of females in the three classes will occur if the initial values are 738, 369, 92 in classes 1, 2, 3, respectively. The growth rate will be 1.2 per 3 years. ■

EXAMPLE 4 Vibrating System of Two Masses on Two Springs (Fig. 161)

Mass–spring systems involving several masses and springs can be treated as eigenvalue problems. For instance, the mechanical system in Fig. 161 is governed by the system of ODEs

$$\begin{aligned} y_1'' &= -3y_1 - 2(y_1 - y_2) = -5y_1 + 2y_2 \\ y_2'' &= \qquad\quad\; -2(y_2 - y_1) = \;\; 2y_1 - 2y_2 \end{aligned} \tag{6}$$

where y_1 and y_2 are the displacements of the masses from rest, as shown in the figure, and primes denote derivatives with respect to time t. In vector form, this becomes

$$y'' = \begin{bmatrix} y_1'' \\ y_2'' \end{bmatrix} = \mathbf{A}\mathbf{y} = \begin{bmatrix} -5 & 2 \\ 2 & -2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \tag{7}$$

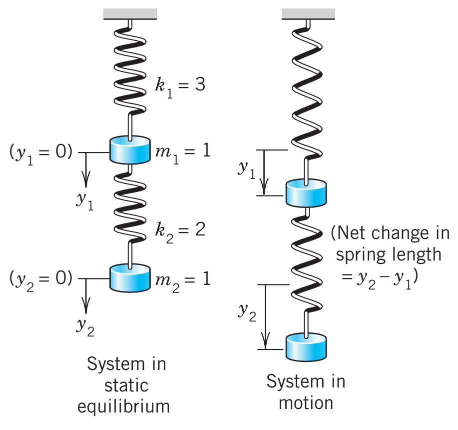


Fig. 161. Masses on springs in Example 4

We try a vector solution of the form

$$\mathbf{y} = \mathbf{x}e^{\omega t}. \tag{8}$$

This is suggested by a mechanical system of a single mass on a spring (Sec. 2.4), whose motion is given by exponential functions (and sines and cosines). Substitution into (7) gives

$$\omega^2 \mathbf{x} e^{\omega t} = A\mathbf{x}e^{\omega t}.$$

Dividing by $e^{\omega t}$ and writing $\omega^2 = \lambda$, we see that our mechanical system leads to the eigenvalue problem

$$\mathbf{A}\mathbf{x} = \lambda \mathbf{x} \qquad \text{where } \lambda = \omega^2. \tag{9}$$

From Example 1 in Sec. 8.1 we see that $\mathbf{A}$ has the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -6$. Consequently, $\omega = \pm\sqrt{-1} = \pm i$ and $\sqrt{-6} = \pm i\sqrt{6}$, respectively. Corresponding eigenvectors are

$$\mathbf{x}_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{x}_2 = \begin{bmatrix} 2 \\ -1 \end{bmatrix}. \tag{10}$$

From (8) we thus obtain the four complex solutions [see (10), Sec. 2.2]

$$\mathbf{x}_1 e^{\pm it} = \mathbf{x}_1(\cos t \pm i \sin t),$$
$$\mathbf{x}_2 e^{\pm i\sqrt{6}t} = \mathbf{x}_2(\cos \sqrt{6}\, t \pm i \sin \sqrt{6}\, t).$$

By addition and subtraction (see Sec. 2.2) we get the four real solutions

$$\mathbf{x}_1 \cos t, \quad \mathbf{x}_1 \sin t, \quad \mathbf{x}_2 \cos \sqrt{6}\, t, \quad \mathbf{x}_2 \sin \sqrt{6}\, t.$$

A general solution is obtained by taking a linear combination of these,

$$\mathbf{y} = \mathbf{x}_1(a_1 \cos t + b_1 \sin t) + \mathbf{x}_2(a_2 \cos \sqrt{6}\, t + b_2 \sin \sqrt{6}\, t)$$

with arbitrary constants a_1, b_1, a_2, b_2 (to which values can be assigned by prescribing initial displacement and initial velocity of each of the two masses). By (10), the components of $\mathbf{y}$ are

$$y_1 = a_1 \cos t + b_1 \sin t + 2a_2 \cos \sqrt{6}\, t + 2b_2 \sin \sqrt{6}\, t$$
$$y_2 = 2a_1 \cos t + 2b_1 \sin t - a_2 \cos \sqrt{6}\, t - b_2 \sin \sqrt{6}\, t.$$

These functions describe harmonic oscillations of the two masses. Physically, this had to be expected because we have neglected damping. ■

PROBLEM SET 8.2

1–6 ELASTIC DEFORMATIONS

Given $\mathbf{A}$ in a deformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, find the principal directions and corresponding factors of extension or contraction. Show the details.

1. $\begin{bmatrix} 3.0 & 1.5 \\ 1.5 & 3.0 \end{bmatrix}$

2. $\begin{bmatrix} 2.0 & 0.4 \\ 0.4 & 2.0 \end{bmatrix}$

3. $\begin{bmatrix} 7 & \sqrt{6} \\ \sqrt{6} & 2 \end{bmatrix}$

4. $\begin{bmatrix} 5 & 2 \\ 2 & 13 \end{bmatrix}$

5. $\begin{bmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{bmatrix}$

6. $\begin{bmatrix} 1.25 & 0.75 \\ 0.75 & 1.25 \end{bmatrix}$

7–9 MARKOV PROCESSES

Find the limit state of the Markov process modeled by the given matrix. Show the details.

7. $\begin{bmatrix} 0.2 & 0.5 \\ 0.8 & 0.5 \end{bmatrix}$

8. $\begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.3 & 0.6 & 0.1 \\ 0.3 & 0.1 & 0.6 \end{bmatrix}$

9. $\begin{bmatrix} 0.6 & 0.1 & 0.2 \\ 0.4 & 0.1 & 0.4 \\ 0 & 0.8 & 0.4 \end{bmatrix}$

10–12 AGE-SPECIFIC POPULATION

Find the growth rate in the Leslie model (see Example 3) with the matrix as given. Show the details.

10. $\begin{bmatrix} 0 & 9.0 & 5.0 \\ 0.4 & 0 & 0 \\ 0 & 0.4 & 0 \end{bmatrix}$ **11.** $\begin{bmatrix} 0 & 3.45 & 0.60 \\ 0.90 & 0 & 0 \\ 0 & 0.45 & 0 \end{bmatrix}$

12. $\begin{bmatrix} 0 & 3.0 & 2.0 & 2.0 \\ 0.5 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0.1 & 0 \end{bmatrix}$

13–15 LEONTIEF MODELS[1]

13. Leontief input–output model. Suppose that three industries are interrelated so that their outputs are used as inputs by themselves, according to the 3×3 **consumption matrix**

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} 0.1 & 0.5 & 0 \\ 0.8 & 0 & 0.4 \\ 0.1 & 0.5 & 0.6 \end{bmatrix}$$

where a_{jk} is the fraction of the output of industry k consumed (purchased) by industry j. Let p_j be the price charged by industry j for its total output. A problem is to find prices so that for each industry, total expenditures equal total income. Show that this leads to $\mathbf{Ap} = \mathbf{p}$, where $\mathbf{p} = [p_1 \quad p_2 \quad p_3]^\mathsf{T}$, and find a solution $\mathbf{p}$ with nonnegative p_1, p_2, p_3.

14. Show that a consumption matrix as considered in Prob. 13 must have column sums 1 and always has the eigenvalue 1.

15. Open Leontief input–output model. If not the whole output but only a portion of it is consumed by the industries themselves, then instead of $\mathbf{Ax} = \mathbf{x}$ (as in Prob. 13), we have $\mathbf{x} - \mathbf{Ax} = \mathbf{y}$, where $\mathbf{x} = [x_1 \quad x_2 \quad x_3]^\mathsf{T}$ is produced, $\mathbf{Ax}$ is consumed by the industries and, thus, $\mathbf{y}$ is the net production available for other consumers. Find for what production $\mathbf{x}$ a given **demand vector** $\mathbf{y} = [0.1 \quad 0.3 \quad 0.1]^\mathsf{T}$ can be achieved if the consumption matrix is

$$\mathbf{A} = \begin{bmatrix} 0.1 & 0.4 & 0.2 \\ 0.5 & 0 & 0.1 \\ 0.1 & 0.4 & 0.4 \end{bmatrix}.$$

16–20 GENERAL PROPERTIES OF EIGENVALUE PROBLEMS

Let $\mathbf{A} = [a_{jk}]$ be an $n \times n$ matrix with (not necessarily distinct) eigenvalues $\lambda_1, \cdots, \lambda_n$. Show.

16. Trace. The sum of the main diagonal entries, called the *trace* of $\mathbf{A}$, equals the sum of the eigenvalues of $\mathbf{A}$.

17. "Spectral shift." $\mathbf{A} - k\mathbf{I}$ has the eigenvalues $\lambda_1 - k, \cdots, \lambda_n - k$ and the same eigenvectors as $\mathbf{A}$.

18. Scalar multiples, powers. $k\mathbf{A}$ has the eigenvalues $k\lambda_1, \cdots, k\lambda_n$. $\mathbf{A}^m (m = 1, 2, \cdots)$ has the eigenvalues $\lambda_1^m, \cdots, \lambda_n^m$. The eigenvectors are those of $\mathbf{A}$.

19. Spectral mapping theorem. The "**polynomial matrix**"

$$p(\mathbf{A}) = k_m\mathbf{A}^m + k_{m-1}\mathbf{A}^{m-1} + \cdots + k_1\mathbf{A} + k_0\mathbf{I}$$

has the eigenvalues

$$p(\lambda_j) = k_m\lambda_j^m + k_{m-1}\lambda_j^{m-1} + \cdots + k_1\lambda_j + k_0$$

where $j = 1, \cdots, n$, and the same eigenvectors as $\mathbf{A}$.

20. Perron's theorem. A Leslie matrix $\mathbf{L}$ with positive $l_{12}, l_{13}, l_{21}, l_{32}$ has a positive eigenvalue. (This is a special case of the Perron–Frobenius theorem in Sec. 20.7, which is difficult to prove in its general form.)

8.3 Symmetric, Skew-Symmetric, and Orthogonal Matrices

We consider three classes of real square matrices that, because of their remarkable properties, occur quite frequently in applications. The first two matrices have already been mentioned in Sec. 7.2. The goal of Sec. 8.3 is to show their remarkable properties.

[1] WASSILY LEONTIEF (1906–1999). American economist at New York University. For his input–output analysis he was awarded the Nobel Prize in 1973.

DEFINITIONS

Symmetric, Skew-Symmetric, and Orthogonal Matrices

A ***real*** square matrix $\mathbf{A} = [a_{jk}]$ is called

symmetric if transposition leaves it unchanged,

$$\mathbf{A}^\mathsf{T} = \mathbf{A}, \qquad \text{thus} \qquad a_{kj} = a_{jk}, \tag{1}$$

skew-symmetric if transposition gives the negative of **A**,

$$\mathbf{A}^\mathsf{T} = -\mathbf{A}, \qquad \text{thus} \qquad a_{kj} = -a_{jk}, \tag{2}$$

orthogonal if transposition gives the inverse of **A**,

$$\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}. \tag{3}$$

EXAMPLE 1 **Symmetric, Skew-Symmetric, and Orthogonal Matrices**

The matrices

$$\begin{bmatrix} -3 & 1 & 5 \\ 1 & 0 & -2 \\ 5 & -2 & 4 \end{bmatrix}, \quad \begin{bmatrix} 0 & 9 & -12 \\ -9 & 0 & 20 \\ 12 & -20 & 0 \end{bmatrix}, \quad \begin{bmatrix} \frac{2}{3} & \frac{1}{3} & \frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} & \frac{1}{3} \\ \frac{1}{3} & \frac{2}{3} & -\frac{2}{3} \end{bmatrix}$$

are symmetric, skew-symmetric, and orthogonal, respectively, as you should verify. Every skew-symmetric matrix has all main diagonal entries zero. (Can you prove this?) ■

Any real square matrix **A** may be written as the sum of a symmetric matrix **R** and a skew-symmetric matrix **S**, where

$$\mathbf{R} = \tfrac{1}{2}(\mathbf{A} + \mathbf{A}^\mathsf{T}) \qquad \text{and} \qquad \mathbf{S} = \tfrac{1}{2}(\mathbf{A} - \mathbf{A}^\mathsf{T}). \tag{4}$$

EXAMPLE 2 **Illustration of Formula (4)**

$$\mathbf{A} = \begin{bmatrix} 9 & 5 & 2 \\ 2 & 3 & -8 \\ 5 & 4 & 3 \end{bmatrix} = \mathbf{R} + \mathbf{S} = \begin{bmatrix} 9.0 & 3.5 & 3.5 \\ 3.5 & 3.0 & -2.0 \\ 3.5 & -2.0 & 3.0 \end{bmatrix} + \begin{bmatrix} 0 & 1.5 & -1.5 \\ -1.5 & 0 & -6.0 \\ 1.5 & 6.0 & 0 \end{bmatrix}$$ ■

THEOREM 1

Eigenvalues of Symmetric and Skew-Symmetric Matrices

(a) *The eigenvalues of a symmetric matrix are real.*

(b) *The eigenvalues of a skew-symmetric matrix are pure imaginary or zero.*

This basic theorem (and an extension of it) will be proved in Sec. 8.5.

EXAMPLE 3 Eigenvalues of Symmetric and Skew-Symmetric Matrices

The matrices in (1) and (7) of Sec. 8.2 are symmetric and have real eigenvalues. The skew-symmetric matrix in Example 1 has the eigenvalues 0, $-25i$, and $25i$. (Verify this.) The following matrix has the real eigenvalues 1 and 5 but is not symmetric. Does this contradict Theorem 1?

$$\begin{bmatrix} 3 & 4 \\ 1 & 3 \end{bmatrix}$$

Orthogonal Transformations and Orthogonal Matrices

Orthogonal transformations are transformations

$$\mathbf{y} = \mathbf{A}\mathbf{x} \qquad \text{where } \mathbf{A} \text{ is an orthogonal matrix.} \tag{5}$$

With each vector $\mathbf{x}$ in R^n such a transformation assigns a vector $\mathbf{y}$ in R^n. For instance, the plane rotation through an angle θ

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \tag{6}$$

is an orthogonal transformation. It can be shown that any orthogonal transformation in the plane or in three-dimensional space is a **rotation** (possibly combined with a reflection in a straight line or a plane, respectively).

The main reason for the importance of orthogonal matrices is as follows.

THEOREM 2 Invariance of Inner Product

An orthogonal transformation preserves the value of the **inner product** *of vectors* $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *defined by*

$$\mathbf{a} \cdot \mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{b} = [a_1 \quad \cdots \quad a_n] \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}. \tag{7}$$

That is, for any $\mathbf{a}$ *and* $\mathbf{b}$ *in* R^n, *orthogonal* $n \times n$ *matrix* $\mathbf{A}$, *and* $\mathbf{u} = \mathbf{A}\mathbf{a}, \mathbf{v} = \mathbf{A}\mathbf{b}$ *we have* $\mathbf{u} \cdot \mathbf{v} = \mathbf{a} \cdot \mathbf{b}$.

Hence the transformation also preserves the **length** *or* **norm** *of any vector* $\mathbf{a}$ *in* R^n *given by*

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{\mathbf{a}^{\mathsf{T}}\mathbf{a}}. \tag{8}$$

PROOF Let $\mathbf{A}$ be orthogonal. Let $\mathbf{u} = \mathbf{A}\mathbf{a}$ and $\mathbf{v} = \mathbf{A}\mathbf{b}$. We must show that $\mathbf{u} \cdot \mathbf{v} = \mathbf{a} \cdot \mathbf{b}$. Now $(\mathbf{A}\mathbf{a})^{\mathsf{T}} = \mathbf{a}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}$ by (10d) in Sec. 7.2 and $\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{I}$ by (3). Hence

$$\mathbf{u} \cdot \mathbf{v} = \mathbf{u}^{\mathsf{T}}\mathbf{v} = (\mathbf{A}\mathbf{a})^{\mathsf{T}}\mathbf{A}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\mathbf{A}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{I}\mathbf{b} = \mathbf{a}^{\mathsf{T}}\mathbf{b} = \mathbf{a} \cdot \mathbf{b}. \tag{9}$$

From this the invariance of $\|\mathbf{a}\|$ follows if we set $\mathbf{b} = \mathbf{a}$. ■

Orthogonal matrices have further interesting properties as follows.

THEOREM 3

Orthonormality of Column and Row Vectors

A real square matrix is orthogonal if and only if its column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$ *(and also its row vectors) form an* **orthonormal system**, *that is,*

$$\mathbf{a}_j \cdot \mathbf{a}_k = \mathbf{a}_j^{\mathsf{T}} \mathbf{a}_k = \begin{cases} 0 & \text{if} \quad j \neq k \\ 1 & \text{if} \quad j = k. \end{cases} \tag{10}$$

PROOF **(a)** Let $\mathbf{A}$ be orthogonal. Then $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}$. In terms of column vectors $\mathbf{a}_1, \cdots, \mathbf{a}_n$,

$$\mathbf{I} = \mathbf{A}^{-1}\mathbf{A} = \mathbf{A}^{\mathsf{T}}\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^{\mathsf{T}} \\ \vdots \\ \mathbf{a}_n^{\mathsf{T}} \end{bmatrix} [\mathbf{a}_1 \cdots \mathbf{a}_n] = \begin{bmatrix} \mathbf{a}_1^{\mathsf{T}}\mathbf{a}_1 & \mathbf{a}_1^{\mathsf{T}}\mathbf{a}_2 & \cdots & \mathbf{a}_1^{\mathsf{T}}\mathbf{a}_n \\ \cdot & \cdot & \cdots & \cdot \\ \mathbf{a}_n^{\mathsf{T}}\mathbf{a}_1 & \mathbf{a}_n^{\mathsf{T}}\mathbf{a}_2 & \cdots & \mathbf{a}_n^{\mathsf{T}}\mathbf{a}_n \end{bmatrix}. \tag{11}$$

The last equality implies (10), by the definition of the $n \times n$ unit matrix $\mathbf{I}$. From (3) it follows that the inverse of an orthogonal matrix is orthogonal (see CAS Experiment 12). Now the column vectors of $\mathbf{A}^{-1}(=\mathbf{A}^{\mathsf{T}})$ are the row vectors of $\mathbf{A}$. Hence the row vectors of $\mathbf{A}$ also form an orthonormal system.

(b) Conversely, if the column vectors of $\mathbf{A}$ satisfy (10), the off-diagonal entries in (11) must be 0 and the diagonal entries 1. Hence $\mathbf{A}^{\mathsf{T}}\mathbf{A} = \mathbf{I}$, as (11) shows. Similarly, $\mathbf{A}\mathbf{A}^{\mathsf{T}} = \mathbf{I}$. This implies $\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$ because also $\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$ and the inverse is unique. Hence $\mathbf{A}$ is orthogonal. Similarly when the row vectors of $\mathbf{A}$ form an orthonormal system, by what has been said at the end of part (**a**). ■

THEOREM 4

Determinant of an Orthogonal Matrix

The determinant of an orthogonal matrix has the value $+1$ *or* -1.

PROOF From $\det \mathbf{AB} = \det \mathbf{A} \det \mathbf{B}$ (Sec. 7.8, Theorem 4) and $\det \mathbf{A}^{\mathsf{T}} = \det \mathbf{A}$ (Sec. 7.7, Theorem 2d), we get for an orthogonal matrix

$$1 = \det \mathbf{I} = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{A}\mathbf{A}^{\mathsf{T}}) = \det \mathbf{A} \det \mathbf{A}^{\mathsf{T}} = (\det \mathbf{A})^2.$$ ■

EXAMPLE 4 **Illustration of Theorems 3 and 4**

The last matrix in Example 1 and the matrix in (6) illustrate Theorems 3 and 4 because their determinants are -1 and $+1$, as you should verify. ■

THEOREM 5

Eigenvalues of an Orthogonal Matrix

The eigenvalues of an orthogonal matrix $\mathbf{A}$ *are real or complex conjugates in pairs and have absolute value* 1.

PROOF The first part of the statement holds for any real matrix **A** because its characteristic polynomial has real coefficients, so that its zeros (the eigenvalues of **A**) must be as indicated. The claim that $|\lambda| = 1$ will be proved in Sec. 8.5. ■

EXAMPLE 5 **Eigenvalues of an Orthogonal Matrix**

The orthogonal matrix in Example 1 has the characteristic equation

$$-\lambda^3 + \tfrac{2}{3}\lambda^2 + \tfrac{2}{3}\lambda - 1 = 0.$$

Now one of the eigenvalues must be real (why?), hence $+1$ or -1. Trying, we find -1. Division by $\lambda + 1$ gives $-(\lambda^2 - 5\lambda/3 + 1) = 0$ and the two eigenvalues $(5 + i\sqrt{11})/6$ and $(5 - i\sqrt{11})/6$, which have absolute value 1. Verify all of this. ■

Looking back at this section, you will find that the numerous basic results it contains have relatively short, straightforward proofs. This is typical of large portions of matrix eigenvalue theory.

PROBLEM SET 8.3

1–10 SPECTRUM

Are the following matrices symmetric, skew-symmetric, or orthogonal? Find the spectrum of each, thereby illustrating Theorems 1 and 5. Show your work in detail.

1. $\begin{bmatrix} 0.8 & 0.6 \\ -0.6 & 0.8 \end{bmatrix}$ **2.** $\begin{bmatrix} a & b \\ -b & a \end{bmatrix}$

3. $\begin{bmatrix} 2 & 8 \\ -8 & 2 \end{bmatrix}$ **4.** $\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$

5. $\begin{bmatrix} 6 & 0 & 0 \\ 0 & 2 & -2 \\ 0 & -2 & 5 \end{bmatrix}$ **6.** $\begin{bmatrix} a & k & k \\ k & a & k \\ k & k & a \end{bmatrix}$

7. $\begin{bmatrix} 0 & 9 & -12 \\ -9 & 0 & 20 \\ 12 & -20 & 0 \end{bmatrix}$ **8.** $\begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix}$

9. $\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ -1 & 0 & 0 \end{bmatrix}$ **10.** $\begin{bmatrix} \frac{4}{9} & \frac{8}{9} & \frac{1}{9} \\ -\frac{7}{9} & \frac{4}{9} & -\frac{4}{9} \\ -\frac{4}{9} & \frac{1}{9} & \frac{8}{9} \end{bmatrix}$

11. WRITING PROJECT. Section Summary. Summarize the main concepts and facts in this section, giving illustrative examples of your own.

12. CAS EXPERIMENT. Orthogonal Matrices.

(a) Products. Inverse. Prove that the product of two orthogonal matrices is orthogonal, and so is the inverse of an orthogonal matrix. What does this mean in terms of rotations?

(b) Rotation. Show that (6) is an orthogonal transformation. Verify that it satisfies Theorem 3. Find the inverse transformation.

(c) Powers. Write a program for computing powers $\mathbf{A}^m$ $(m = 1, 2, \cdots)$ of a 2×2 matrix **A** and their spectra. Apply it to the matrix in Prob. 1 (call it **A**). To what rotation does **A** correspond? Do the eigenvalues of $\mathbf{A}^m$ have a limit as $m \to \infty$?

(d) Compute the eigenvalues of $(0.9\mathbf{A})^m$, where **A** is the matrix in Prob. 1. Plot them as points. What is their limit? Along what kind of curve do these points approach the limit?

(e) Find **A** such that $\mathbf{y} = \mathbf{A}\mathbf{x}$ is a counterclockwise rotation through 30° in the plane.

13–20 GENERAL PROPERTIES

13. Verification. Verify the statements in Example 1.

14. Verify the statements in Examples 3 and 4.

15. Sum. Are the eigenvalues of $\mathbf{A} + \mathbf{B}$ sums of the eigenvalues of **A** and of **B**?

16. Orthogonality. Prove that eigenvectors of a symmetric matrix corresponding to different eigenvalues are orthogonal. Give examples.

17. Skew-symmetric matrix. Show that the inverse of a skew-symmetric matrix is skew-symmetric.

18. Do there exist nonsingular skew-symmetric $n \times n$ matrices with odd n?

19. Orthogonal matrix. Do there exist skew-symmetric orthogonal 3×3 matrices?

20. Symmetric matrix. Do there exist nondiagonal symmetric 3×3 matrices that are orthogonal?

8.4 Eigenbases. Diagonalization. Quadratic Forms

So far we have emphasized properties of eigen*values*. We now turn to general properties of eigen*vectors*. Eigenvectors of an $n \times n$ matrix $\mathbf{A}$ may (or may not!) form a basis for R^n. If we are interested in a transformation $\mathbf{y} = \mathbf{A}\mathbf{x}$, such an "**eigenbasis**" (basis of eigenvectors)—if it exists—is of great advantage because then we can represent any $\mathbf{x}$ in R^n uniquely as a linear combination of the eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$, say,

$$\mathbf{x} = c_1\mathbf{x}_1 + c_2\mathbf{x}_2 + \cdots + c_n\mathbf{x}_n.$$

And, denoting the corresponding (not necessarily distinct) eigenvalues of the matrix $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_n$, we have $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j$, so that we simply obtain

$$\begin{aligned} \mathbf{y} = \mathbf{A}\mathbf{x} &= \mathbf{A}(c_1\mathbf{x}_1 + \cdots + c_n\mathbf{x}_n) \\ &= c_1\mathbf{A}\mathbf{x}_1 + \cdots + c_n\mathbf{A}\mathbf{x}_n \\ &= c_1\lambda_1\mathbf{x}_1 + \cdots + c_n\lambda_n\mathbf{x}_n. \end{aligned} \tag{1}$$

This shows that we have decomposed the complicated action of $\mathbf{A}$ on an arbitrary vector $\mathbf{x}$ into a sum of simple actions (multiplication by scalars) on the eigenvectors of $\mathbf{A}$. This is the point of an eigenbasis.

Now if the n eigenvalues are all different, we do obtain a basis:

THEOREM 1

Basis of Eigenvectors

*If an $n \times n$ matrix $\mathbf{A}$ has n **distinct** eigenvalues, then $\mathbf{A}$ has a basis of eigenvectors $\mathbf{x}_1, \cdots, \mathbf{x}_n$ for R^n.*

PROOF All we have to show is that $\mathbf{x}_1, \cdots, \mathbf{x}_n$ are linearly independent. Suppose they are not. Let r be the largest integer such that $\{\mathbf{x}_1, \cdots, \mathbf{x}_r\}$ is a linearly independent set. Then $r < n$ and the set $\{\mathbf{x}_1, \cdots, \mathbf{x}_r, \mathbf{x}_{r+1}\}$ is linearly dependent. Thus there are scalars $c_1, \cdots, c_{r+1}$, not all zero, such that

$$c_1\mathbf{x}_1 + \cdots + c_{r+1}\mathbf{x}_{r+1} = \mathbf{0} \tag{2}$$

(see Sec. 7.4). Multiplying both sides by $\mathbf{A}$ and using $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j$, we obtain

$$\mathbf{A}(c_1\mathbf{x}_1 + \cdots + c_{r+1}\mathbf{x}_{r+1}) = c_1\lambda_1\mathbf{x}_1 + \cdots + c_{r+1}\lambda_{r+1}\mathbf{x}_{r+1} = \mathbf{A}\mathbf{0} = \mathbf{0}. \tag{3}$$

To get rid of the last term, we subtract λ_{r+1} times (2) from this, obtaining

$$c_1(\lambda_1 - \lambda_{r+1})\mathbf{x}_1 + \cdots + c_r(\lambda_r - \lambda_{r+1})\mathbf{x}_r = \mathbf{0}.$$

Here $c_1(\lambda_1 - \lambda_{r+1}) = 0, \cdots, c_r(\lambda_r - \lambda_{r+1}) = 0$ since $\{x_1, \cdots, x_r\}$ is linearly independent. Hence $c_1 = \cdots = c_r = 0$, since all the eigenvalues are distinct. But with this, (2) reduces to $c_{r+1}\mathbf{x}_{r+1} = \mathbf{0}$, hence $c_{r+1} = 0$, since $\mathbf{x}_{r+1} \neq \mathbf{0}$ (an eigenvector!). This contradicts the fact that not all scalars in (2) are zero. Hence the conclusion of the theorem must hold. ■

EXAMPLE 1 **Eigenbasis. Nondistinct Eigenvalues. Nonexistence**

The matrix $\mathbf{A} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$ has a basis of eigenvectors $\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ corresponding to the eigenvalues $\lambda_1 = 8$, $\lambda_2 = 2$. (See Example 1 in Sec. 8.2.)

Even if not all n eigenvalues are different, a matrix $\mathbf{A}$ may still provide an eigenbasis for R^n. See Example 2 in Sec. 8.1, where $n = 3$.

On the other hand, $\mathbf{A}$ ***may not have enough linearly independent eigenvectors to make up a basis***. For instance, $\mathbf{A}$ in Example 3 of Sec. 8.1 is

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} \quad \text{and has only one eigenvector} \quad \begin{bmatrix} k \\ 0 \end{bmatrix} \quad (k \neq 0, \text{ arbitrary}).$$ ■

Actually, eigenbases exist under much more general conditions than those in Theorem 1. An important case is the following.

THEOREM 2 **Symmetric Matrices**

A symmetric matrix has an orthonormal basis of eigenvectors for R^n.

For a proof (which is involved) see Ref. [B3], vol. 1, pp. 270–272.

EXAMPLE 2 **Orthonormal Basis of Eigenvectors**

The first matrix in Example 1 is symmetric, and an orthonormal basis of eigenvectors is $[1/\sqrt{2} \quad 1/\sqrt{2}]^\mathsf{T}$, $[1/\sqrt{2} \quad -1/\sqrt{2}]^\mathsf{T}$. ■

Similarity of Matrices. Diagonalization

Eigenbases also play a role in reducing a matrix $\mathbf{A}$ to a diagonal matrix whose entries are the eigenvalues of $\mathbf{A}$. This is done by a "similarity transformation," which is defined as follows (and will have various applications in numerics in Chap. 20).

DEFINITION **Similar Matrices. Similarity Transformation**

An $n \times n$ matrix $\hat{\mathbf{A}}$ is called **similar** to an $n \times n$ matrix $\mathbf{A}$ if

$$\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \tag{4}$$

for some (nonsingular!) $n \times n$ matrix $\mathbf{P}$. This transformation, which gives $\hat{\mathbf{A}}$ from $\mathbf{A}$, is called a **similarity transformation**.

The key property of this transformation is that it preserves the eigenvalues of $\mathbf{A}$:

THEOREM 3 **Eigenvalues and Eigenvectors of Similar Matrices**

If $\hat{\mathbf{A}}$ is similar to $\mathbf{A}$, then $\hat{\mathbf{A}}$ has the same eigenvalues as $\mathbf{A}$.

Furthermore, if $\mathbf{x}$ is an eigenvector of $\mathbf{A}$, then $\mathbf{y} = \mathbf{P}^{-1}\mathbf{x}$ is an eigenvector of $\hat{\mathbf{A}}$ corresponding to the same eigenvalue.

PROOF From $\mathbf{Ax} = \lambda\mathbf{x}$ (λ an eigenvalue, $\mathbf{x} \neq \mathbf{0}$) we get $\mathbf{P}^{-1}\mathbf{Ax} = \lambda\mathbf{P}^{-1}\mathbf{x}$. Now $\mathbf{I} = \mathbf{PP}^{-1}$. By this ***identity trick*** the equation $\mathbf{P}^{-1}\mathbf{Ax} = \lambda\mathbf{P}^{-1}\mathbf{x}$ gives

$$\mathbf{P}^{-1}\mathbf{Ax} = \mathbf{P}^{-1}\mathbf{AIx} = \mathbf{P}^{-1}\mathbf{APP}^{-1}\mathbf{x} = (\mathbf{P}^{-1}\mathbf{AP})\mathbf{P}^{-1}\mathbf{x} = \hat{\mathbf{A}}(\mathbf{P}^{-1}\mathbf{x}) = \lambda\mathbf{P}^{-1}\mathbf{x}.$$

Hence λ is an eigenvalue of $\hat{\mathbf{A}}$ and $\mathbf{P}^{-1}\mathbf{x}$ a corresponding eigenvector. Indeed, $\mathbf{P}^{-1}\mathbf{x} \neq \mathbf{0}$ because $\mathbf{P}^{-1}\mathbf{x} = \mathbf{0}$ would give $\mathbf{x} = \mathbf{Ix} = \mathbf{PP}^{-1}\mathbf{x} = \mathbf{P0} = \mathbf{0}$, contradicting $\mathbf{x} \neq \mathbf{0}$. ■

EXAMPLE 3 **Eigenvalues and Vectors of Similar Matrices**

Let,
$$\mathbf{A} = \begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix} \quad \text{and} \quad \mathbf{P} = \begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix}.$$

Then
$$\hat{\mathbf{A}} = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 6 & -3 \\ 4 & -1 \end{bmatrix}\begin{bmatrix} 1 & 3 \\ 1 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 0 \\ 0 & 2 \end{bmatrix}.$$

Here $\mathbf{P}^{-1}$ was obtained from (4*) in Sec. 7.8 with det $\mathbf{P} = 1$. We see that $\hat{\mathbf{A}}$ has the eigenvalues $\lambda_1 = 3$, $\lambda_2 = 2$. The characteristic equation of $\mathbf{A}$ is $(6 - \lambda)(-1 - \lambda) + 12 = \lambda^2 - 5\lambda + 6 = 0$. It has the roots (the eigenvalues of $\mathbf{A}$) $\lambda_1 = 3$, $\lambda_2 = 2$, confirming the first part of Theorem 3.

We confirm the second part. From the first component of $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ we have $(6 - \lambda)x_1 - 3x_2 = 0$. For $\lambda = 3$ this gives $3x_1 - 3x_2 = 0$, say, $\mathbf{x}_1 = [1 \quad 1]^\mathsf{T}$. For $\lambda = 2$ it gives $4x_1 - 3x_2 = 0$, say, $\mathbf{x}_2 = [3 \quad 4]^\mathsf{T}$. In Theorem 3 we thus have

$$\mathbf{y}_1 = \mathbf{P}^{-1}\mathbf{x}_1 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{y}_2 = \mathbf{P}^{-1}\mathbf{x}_2 = \begin{bmatrix} 4 & -3 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 3 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Indeed, these are eigenvectors of the diagonal matrix $\hat{\mathbf{A}}$.

Perhaps we see that $\mathbf{x}_1$ and $\mathbf{x}_2$ are the columns of $\mathbf{P}$. This suggests the general method of transforming a matrix $\mathbf{A}$ to diagonal form $\mathbf{D}$ by using $\mathbf{P} = \mathbf{X}$, the matrix with eigenvectors as columns. ■

By a suitable similarity transformation we can now transform a matrix $\mathbf{A}$ to a diagonal matrix $\mathbf{D}$ whose diagonal entries are the eigenvalues of $\mathbf{A}$:

THEOREM 4 **Diagonalization of a Matrix**

If an $n \times n$ matrix $\mathbf{A}$ has a basis of eigenvectors, then

(5) $$\mathbf{D} = \mathbf{X}^{-1}\mathbf{AX}$$

is diagonal, with the eigenvalues of $\mathbf{A}$ as the entries on the main diagonal. Here $\mathbf{X}$ is the matrix with these eigenvectors as column vectors. Also,

(5*) $$\mathbf{D}^m = \mathbf{X}^{-1}\mathbf{A}^m\mathbf{X} \qquad (m = 2, 3, \cdots).$$

PROOF Let $\mathbf{x}_1, \cdots, \mathbf{x}_n$ be a basis of eigenvectors of $\mathbf{A}$ for R^n. Let the corresponding eigenvalues of $\mathbf{A}$ be $\lambda_1, \cdots, \lambda_n$, respectively, so that $\mathbf{A}\mathbf{x}_1 = \lambda_1\mathbf{x}_1, \cdots, \mathbf{A}\mathbf{x}_n = \lambda_n\mathbf{x}_n$. Then $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_n]$ has rank n, by Theorem 3 in Sec. 7.4. Hence $\mathbf{X}^{-1}$ exists by Theorem 1 in Sec. 7.8. We claim that

$$(6) \qquad \mathbf{A}\mathbf{x} = \mathbf{A}[\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_n] = [\mathbf{A}\mathbf{x}_1 \quad \cdots \quad \mathbf{A}\mathbf{x}_n] = [\lambda_1\mathbf{x}_1 \quad \cdots \quad \lambda_n\mathbf{x}_n] = \mathbf{X}\mathbf{D}$$

where $\mathbf{D}$ is the diagonal matrix as in (5). The fourth equality in (6) follows by direct calculation. (Try it for $n = 2$ and then for general n.) The third equality uses $\mathbf{A}\mathbf{x}_k = \lambda_k\mathbf{x}_k$. The second equality results if we note that the first column of $\mathbf{AX}$ is $\mathbf{A}$ times the first column of $\mathbf{X}$, which is $\mathbf{x}_1$, and so on. For instance, when $n = 2$ and we write $\mathbf{x}_1 = [x_{11} \quad x_{21}]$, $\mathbf{x}_2 = [x_{12} \quad x_{22}]$, we have

$$\mathbf{AX} = \mathbf{A}[\mathbf{x}_1 \quad \mathbf{x}_2] = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}$$

$$= \underset{\text{Column 1} \qquad\qquad\qquad \text{Column 2}}{\begin{bmatrix} a_{11}x_{11} + a_{12}x_{21} & a_{11}x_{12} + a_{12}x_{22} \\ a_{21}x_{11} + a_{22}x_{21} & a_{21}x_{12} + a_{22}x_{22} \end{bmatrix}} = [\mathbf{A}\mathbf{x}_1 \quad \mathbf{A}\mathbf{x}_2].$$

If we multiply (6) by $\mathbf{X}^{-1}$ from the left, we obtain (5). Since (5) is a similarity transformation, Theorem 3 implies that $\mathbf{D}$ has the same eigenvalues as $\mathbf{A}$. Equation (5*) follows if we note that

$$\mathbf{D}^2 = \mathbf{DD} = (\mathbf{X}^{-1}\mathbf{AX})(\mathbf{X}^{-1}\mathbf{AX}) = \mathbf{X}^{-1}\mathbf{A}(\mathbf{XX}^{-1})\mathbf{AX} = \mathbf{X}^{-1}\mathbf{AAX} = \mathbf{X}^{-1}\mathbf{A}^2\mathbf{X}, \quad \text{etc.} \ \blacksquare$$

EXAMPLE 4 **Diagonalization**

Diagonalize

$$\mathbf{A} = \begin{bmatrix} 7.3 & 0.2 & -3.7 \\ -11.5 & 1.0 & 5.5 \\ 17.7 & 1.8 & -9.3 \end{bmatrix}.$$

Solution. The characteristic determinant gives the characteristic equation $-\lambda^3 - \lambda^2 + 12\lambda = 0$. The roots (eigenvalues of $\mathbf{A}$) are $\lambda_1 = 3$, $\lambda_2 = -4$, $\lambda_3 = 0$. By the Gauss elimination applied to $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1, \lambda_2, \lambda_3$ we find eigenvectors and then $\mathbf{X}^{-1}$ by the Gauss–Jordan elimination (Sec. 7.8, Example 1). The results are

$$\begin{bmatrix} -1 \\ 3 \\ -1 \end{bmatrix}, \quad \begin{bmatrix} 1 \\ -1 \\ 3 \end{bmatrix}, \quad \begin{bmatrix} 2 \\ 1 \\ 4 \end{bmatrix}, \qquad \mathbf{X} = \begin{bmatrix} -1 & 1 & 2 \\ 3 & -1 & 1 \\ -1 & 3 & 4 \end{bmatrix}, \qquad \mathbf{X}^{-1} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix}.$$

Calculating $\mathbf{AX}$ and multiplying by $\mathbf{X}^{-1}$ from the left, we thus obtain

$$\mathbf{D} = \mathbf{X}^{-1}\mathbf{AX} = \begin{bmatrix} -0.7 & 0.2 & 0.3 \\ -1.3 & -0.2 & 0.7 \\ 0.8 & 0.2 & -0.2 \end{bmatrix} \begin{bmatrix} -3 & -4 & 0 \\ 9 & 4 & 0 \\ -3 & -12 & 0 \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 0 \end{bmatrix}. \quad \blacksquare$$

Quadratic Forms. Transformation to Principal Axes

By definition, a **quadratic form** Q in the components $x_1, \cdots, x_n$ of a vector $\mathbf{x}$ is a sum of n^2 terms, namely,

$$
\textbf{(7)} \qquad
\begin{aligned}
Q = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k \\
&= \quad a_{11}x_1^2 \quad + a_{12}x_1x_2 + \cdots + a_{1n}x_1x_n \\
&\quad + a_{21}x_2x_1 + a_{22}x_2^2 \quad + \cdots + a_{2n}x_2x_n \\
&\quad + \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
&\quad + a_{n1}x_nx_1 + a_{n2}x_nx_2 + \cdots + a_{nn}x_n^2.
\end{aligned}
$$

$\mathbf{A} = [a_{jk}]$ is called the **coefficient matrix** of the form. We may assume that $\mathbf{A}$ is ***symmetric***, because we can take off-diagonal terms together in pairs and write the result as a sum of two equal terms; see the following example.

EXAMPLE 5 Quadratic Form. Symmetric Coefficient Matrix

Let

$$\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 4 \\ 6 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 4x_1x_2 + 6x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2.$$

Here $4 + 6 = 10 = 5 + 5$. From the corresponding ***symmetric*** matrix $\mathbf{C} = [c_{jk}]$, where $c_{jk} = \frac{1}{2}(a_{jk} + a_{kj})$, thus $c_{11} = 3$, $c_{12} = c_{21} = 5$, $c_{22} = 2$, we get the same result; indeed,

$$\mathbf{x}^\mathsf{T}\mathbf{C}\mathbf{x} = [x_1 \quad x_2]\begin{bmatrix} 3 & 5 \\ 5 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 3x_1^2 + 5x_1x_2 + 5x_2x_1 + 2x_2^2 = 3x_1^2 + 10x_1x_2 + 2x_2^2.$$ ■

Quadratic forms occur in physics and geometry, for instance, in connection with conic sections (ellipses $x_1^2/a^2 + x_2^2/b^2 = 1$, etc.) and quadratic surfaces (cones, etc.). Their transformation to principal axes is an important practical task related to the diagonalization of matrices, as follows.

By Theorem 2, the *symmetric* coefficient matrix $\mathbf{A}$ of (7) has an orthonormal basis of eigenvectors. Hence if we take these as column vectors, we obtain a matrix $\mathbf{X}$ that is orthogonal, so that $\mathbf{X}^{-1} = \mathbf{X}^\mathsf{T}$. From (5) we thus have $\mathbf{A} = \mathbf{X}\mathbf{D}\mathbf{X}^{-1} = \mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}$. Substitution into (7) gives

$$(8) \qquad Q = \mathbf{x}^\mathsf{T}\mathbf{X}\mathbf{D}\mathbf{X}^\mathsf{T}\mathbf{x}.$$

If we set $\mathbf{X}^\mathsf{T}\mathbf{x} = \mathbf{y}$, then, since $\mathbf{X}^\mathsf{T} = \mathbf{X}^{-1}$, we have $\mathbf{X}^{-1}\mathbf{x} = \mathbf{y}$ and thus obtain

$$(9) \qquad \mathbf{x} = \mathbf{X}\mathbf{y}.$$

Furthermore, in (8) we have $\mathbf{x}^\mathsf{T}\mathbf{X} = (\mathbf{X}^\mathsf{T}\mathbf{x})^\mathsf{T} = \mathbf{y}^\mathsf{T}$ and $\mathbf{X}^\mathsf{T}\mathbf{x} = \mathbf{y}$, so that Q becomes simply

$$\textbf{(10)} \qquad Q = \mathbf{y}^\mathsf{T}\mathbf{D}\mathbf{y} = \lambda_1 y_1^2 + \lambda_2 y_2^2 + \cdots + \lambda_n y_n^2.$$

This proves the following basic theorem.

THEOREM 5

Principal Axes Theorem

The substitution (9) *transforms a quadratic form*

$$Q = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}x_j x_k \qquad (a_{kj} = a_{jk})$$

to the principal axes form or **canonical form** (10), *where* $\lambda_1, \cdots, \lambda_n$ *are the* (*not necessarily distinct*) *eigenvalues of the* (*symmetric!*) *matrix* $\mathbf{A}$, *and* $\mathbf{X}$ *is an orthogonal matrix with corresponding eigenvectors* $\mathbf{x}_1, \cdots, \mathbf{x}_n$, *respectively, as column vectors.*

EXAMPLE 6

Transformation to Principal Axes. Conic Sections

Find out what type of conic section the following quadratic form represents and transform it to principal axes:

$$Q = 17x_1^2 - 30x_1x_2 + 17x_2^2 = 128.$$

Solution. We have $Q = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$, where

$$\mathbf{A} = \begin{bmatrix} 17 & -15 \\ -15 & 17 \end{bmatrix}, \qquad \mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}.$$

This gives the characteristic equation $(17 - \lambda)^2 - 15^2 = 0$. It has the roots $\lambda_1 = 2$, $\lambda_2 = 32$. Hence (10) becomes

$$Q = 2y_1^2 + 32y_2^2.$$

We see that $Q = 128$ represents the ellipse $2y_1^2 + 32y_2^2 = 128$, that is,

$$\frac{y_1^2}{8^2} + \frac{y_2^2}{2^2} = 1.$$

If we want to know the direction of the principal axes in the x_1x_2-coordinates, we have to determine normalized eigenvectors from $(\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}$ with $\lambda = \lambda_1 = 2$ and $\lambda = \lambda_2 = 32$ and then use (9). We get

$$\begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} -1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix},$$

hence

$$\mathbf{x} = \mathbf{X}\mathbf{y} = \begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}, \qquad \begin{aligned} x_1 &= y_1/\sqrt{2} - y_2/\sqrt{2} \\ x_2 &= y_1/\sqrt{2} + y_2/\sqrt{2}. \end{aligned}$$

This is a 45° rotation. Our results agree with those in Sec. 8.2, Example 1, except for the notations. See also Fig. 160 in that example. ■

PROBLEM SET 8.4

1–5 SIMILAR MATRICES HAVE EQUAL EIGENVALUES

Verify this for $\mathbf{A}$ and $\hat{\mathbf{A}} = \mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. If $\mathbf{y}$ is an eigenvector of $\mathbf{P}$, show that $\mathbf{x} = \mathbf{P}\mathbf{y}$ are eigenvectors of $\mathbf{A}$. Show the details of your work.

1. $\mathbf{A} = \begin{bmatrix} 3 & 4 \\ 4 & -3 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} -4 & 2 \\ 3 & -1 \end{bmatrix}$

2. $\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 7 & -5 \\ 10 & -7 \end{bmatrix}$

3. $\mathbf{A} = \begin{bmatrix} 8 & -4 \\ 2 & 2 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0.28 & 0.96 \\ -0.96 & 0.28 \end{bmatrix}$

4. $\mathbf{A} = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 3 & 2 \\ 1 & 0 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 2 & 0 & 3 \\ 0 & 1 & 0 \\ 3 & 0 & 5 \end{bmatrix},$
$\lambda_1 = 3$

5. $\mathbf{A} = \begin{bmatrix} -5 & 0 & 15 \\ 3 & 4 & -9 \\ -5 & 0 & 15 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

6. **PROJECT. Similarity of Matrices.** Similarity is basic, for instance, in designing numeric methods.

   (a) Trace. By definition, the **trace** of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the sum of the diagonal entries,

   $$\text{trace } \mathbf{A} = a_{11} + a_{22} + \cdots + a_{nn}.$$

   Show that the trace equals the sum of the eigenvalues, each counted as often as its algebraic multiplicity indicates. Illustrate this with the matrices $\mathbf{A}$ in Probs. 1, 3, and 5.

   (b) Trace of product. Let $\mathbf{B} = [b_{jk}]$ be $n \times n$. Show that similar matrices have equal traces, by first proving

   $$\text{trace } \mathbf{AB} = \sum_{i=1}^{n} \sum_{l=1}^{n} a_{il}b_{li} = \text{trace } \mathbf{BA}.$$

   (c) Find a relationship between $\hat{\mathbf{A}}$ in (4) and $\hat{\hat{\mathbf{A}}} = \mathbf{P}\hat{\mathbf{A}}\mathbf{P}^{-1}$.

   (d) Diagonalization. What can you do in (5) if you want to change the order of the eigenvalues in $\mathbf{D}$, for instance, interchange $d_{11} = \lambda_1$ and $d_{22} = \lambda_2$?

7. **No basis.** Find further 2×2 and 3×3 matrices without eigenbasis.

8. **Orthonormal basis.** Illustrate Theorem 2 with further examples.

9–16 DIAGONALIZATION OF MATRICES

Find an eigenbasis (a basis of eigenvectors) and diagonalize. Show the details.

9. $\begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}$

10. $\begin{bmatrix} 1 & 0 \\ 2 & -1 \end{bmatrix}$

11. $\begin{bmatrix} -19 & 7 \\ -42 & 16 \end{bmatrix}$

12. $\begin{bmatrix} -4.3 & 7.7 \\ 1.3 & 9.3 \end{bmatrix}$

13. $\begin{bmatrix} 4 & 0 & 0 \\ 12 & -2 & 0 \\ 21 & -6 & 1 \end{bmatrix}$

14. $\begin{bmatrix} -5 & -6 & 6 \\ -9 & -8 & 12 \\ -12 & -12 & 16 \end{bmatrix}, \quad \lambda_1 = -2$

15. $\begin{bmatrix} 4 & 3 & 3 \\ 3 & 6 & 1 \\ 3 & 1 & 6 \end{bmatrix}, \quad \lambda_1 = 10$

16. $\begin{bmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & -4 \end{bmatrix}$

17–23 PRINCIPAL AXES. CONIC SECTIONS

What kind of conic section (or pair of straight lines) is given by the quadratic form? Transform it to principal axes. Express $\mathbf{x}^\mathsf{T} = [x_1 \;\; x_2]$ in terms of the new coordinate vector $\mathbf{y}^\mathsf{T} = [y_1 \;\; y_2]$, as in Example 6.

17. $7x_1^2 + 6x_1x_2 + 7x_2^2 = 200$

18. $3x_1^2 + 8x_1x_2 - 3x_2^2 = 10$

19. $3x_1^2 + 22x_1x_2 + 3x_2^2 = 0$

20. $9x_1^2 + 6x_1x_2 + x_2^2 = 10$

21. $x_1^2 - 12x_1x_2 + x_2^2 = 70$

22. $4x_1^2 + 12x_1x_2 + 13x_2^2 = 16$

23. $-11x_1^2 + 84x_1x_2 + 24x_2^2 = 156$

24. Definiteness. A quadratic form $Q(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ and its (symmetric!) matrix $\mathbf{A}$ are called **(a) positive definite** if $Q(\mathbf{x}) > 0$ for all $\mathbf{x} \neq \mathbf{0}$, **(b) negative definite** if $Q(\mathbf{x}) < 0$ for all $\mathbf{x} \neq \mathbf{0}$, **(c) indefinite** if $Q(\mathbf{x})$ takes both positive and negative values. (See Fig. 162.) [$Q(\mathbf{x})$ and $\mathbf{A}$ are called *positive semidefinite* (*negative semidefinite*) if $Q(\mathbf{x}) \geqq 0$ ($Q(\mathbf{x}) \leqq 0$) for all $\mathbf{x}$.] Show that a necessary and sufficient condition for (a), (b), and (c) is that the eigenvalues of $\mathbf{A}$ are (a) all positive, (b) all negative, and (c) both positive and negative.

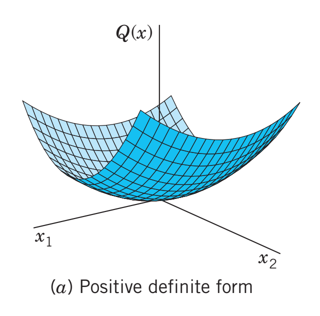
Hint. Use Theorem 5.

25. Definiteness. A necessary and sufficient condition for positive definiteness of a quadratic form $Q(\mathbf{x}) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ with *symmetric* matrix $\mathbf{A}$ is that all the **principal minors** are positive (see Ref. [B3], vol. 1, p. 306), that is,

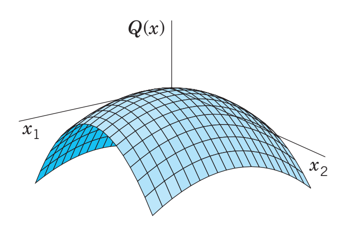
$$a_{11} > 0, \qquad \begin{vmatrix} a_{11} & a_{12} \\ a_{12} & a_{22} \end{vmatrix} > 0,$$

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{vmatrix} > 0, \qquad \cdots, \qquad \det \mathbf{A} > 0.$$

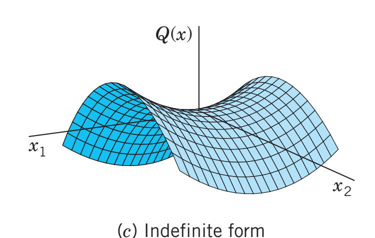
Show that the form in Prob. 22 is positive definite, whereas that in Prob. 23 is indefinite.



(a) Positive definite form



(b) Negative definite form



(c) Indefinite form

Fig. 162. Quadratic forms in two variables (Problem 24)

8.5 Complex Matrices and Forms. *Optional*

The three classes of matrices in Sec. 8.3 have complex counterparts which are of practical interest in certain applications, for instance, in quantum mechanics. This is mainly because of their spectra as shown in Theorem 1 in this section. The second topic is about extending quadratic forms of Sec. 8.4 to complex numbers. (The reader who wants to brush up on complex numbers may want to consult Sec. 13.1.)

Notations

$\overline{\mathbf{A}} = [\overline{a}_{jk}]$ is obtained from $\mathbf{A} = [a_{jk}]$ by replacing each entry $a_{jk} = \alpha + i\beta$ (α, β real) with its complex conjugate $\overline{a}_{jk} = \alpha - i\beta$. Also, $\overline{\mathbf{A}}^{\mathsf{T}} = [\overline{a}_{kj}]$ is the transpose of $\overline{\mathbf{A}}$, hence the conjugate transpose of $\mathbf{A}$.

EXAMPLE 1 Notations

If $\mathbf{A} = \begin{bmatrix} 3+4i & 1-i \\ 6 & 2-5i \end{bmatrix}$, then $\overline{\mathbf{A}} = \begin{bmatrix} 3-4i & 1+i \\ 6 & 2+5i \end{bmatrix}$ and $\overline{\mathbf{A}}^{\mathsf{T}} = \begin{bmatrix} 3-4i & 6 \\ 1+i & 2+5i \end{bmatrix}$. ■

DEFINITION

Hermitian, Skew-Hermitian, and Unitary Matrices

A square matrix $\mathbf{A} = [a_{kj}]$ is called

Hermitian	if	$\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$,	that is,	$\overline{a}_{kj} = a_{jk}$
skew-Hermitian	if	$\overline{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$,	that is,	$\overline{a}_{kj} = -a_{jk}$
unitary	if	$\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$.		

The first two classes are named after Hermite (see footnote 13 in Problem Set 5.8).

From the definitions we see the following. If $\mathbf{A}$ is Hermitian, the entries on the main diagonal must satisfy $\overline{a}_{jj} = a_{jj}$; that is, they are real. Similarly, if $\mathbf{A}$ is skew-Hermitian, then $\overline{a}_{jj} = -a_{jj}$. If we set $a_{jj} = \alpha + i\beta$, this becomes $\alpha - i\beta = -(\alpha + i\beta)$. Hence $\alpha = 0$, so that a_{jj} must be pure imaginary or 0.

EXAMPLE 2 Hermitian, Skew-Hermitian, and Unitary Matrices

$$\mathbf{A} = \begin{bmatrix} 4 & 1-3i \\ 1+3i & 7 \end{bmatrix} \qquad \mathbf{B} = \begin{bmatrix} 3i & 2+i \\ -2+i & -i \end{bmatrix} \qquad \mathbf{C} = \begin{bmatrix} \frac{1}{2}i & \frac{1}{2}\sqrt{3} \\ \frac{1}{2}\sqrt{3} & \frac{1}{2}i \end{bmatrix}$$

are Hermitian, skew-Hermitian, and unitary matrices, respectively, as you may verify by using the definitions. ■

If a Hermitian matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}$. Hence a real Hermitian matrix is a symmetric matrix (Sec. 8.3).

Similarly, if a skew-Hermitian matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = -\mathbf{A}$. Hence a real skew-Hermitian matrix is a skew-symmetric matrix.

Finally, if a unitary matrix is real, then $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$. Hence a real unitary matrix is an orthogonal matrix.

This shows that *Hermitian, skew-Hermitian, and unitary matrices generalize symmetric, skew-symmetric, and orthogonal matrices, respectively.*

Eigenvalues

It is quite remarkable that the matrices under consideration have spectra (sets of eigenvalues; see Sec. 8.1) that can be characterized in a general way as follows (see Fig. 163).

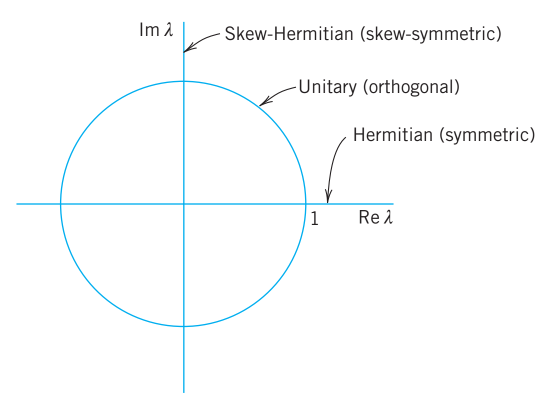


Fig. 163. Location of the eigenvalues of Hermitian, skew-Hermitian, and unitary matrices in the complex λ-plane

THEOREM 1

Eigenvalues

(a) *The eigenvalues of a Hermitian matrix (and thus of a symmetric matrix) are real.*

(b) *The eigenvalues of a skew-Hermitian matrix (and thus of a skew-symmetric matrix) are pure imaginary or zero.*

(c) *The eigenvalues of a unitary matrix (and thus of an orthogonal matrix) have absolute value* 1.

EXAMPLE 3 **Illustration of Theorem 1**

For the matrices in Example 2 we find by direct calculation

	Matrix	Characteristic Equation	Eigenvalues
A	Hermitian	$\lambda^2 - 11\lambda + 18 = 0$	$9, \quad 2$
B	Skew-Hermitian	$\lambda^2 - 2i\lambda + 8 = 0$	$4i, \quad -2i$
C	Unitary	$\lambda^2 - i\lambda - 1 = 0$	$\frac{1}{2}\sqrt{3} + \frac{1}{2}i, \quad -\frac{1}{2}\sqrt{3} + \frac{1}{2}i$

and $|\pm\frac{1}{2}\sqrt{3} + \frac{1}{2}i|^2 = \frac{3}{4} + \frac{1}{4} = 1.$ ■

PROOF We prove Theorem 1. Let λ be an eigenvalue and $\mathbf{x}$ an eigenvector of $\mathbf{A}$. Multiply $\mathbf{Ax} = \lambda\mathbf{x}$ from the left by $\bar{\mathbf{x}}^{\mathsf{T}}$, thus $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax} = \lambda\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$, and divide by $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x} = \bar{x}_1x_1 + \cdots + \bar{x}_nx_n = |x_1|^2 + \cdots + |x_n|^2$, which is real and not 0 because $\mathbf{x} \neq \mathbf{0}$. This gives

$$\lambda = \frac{\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}}{\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}}. \tag{1}$$

(a) If $\mathbf{A}$ is Hermitian, $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$ or $\mathbf{A}^{\mathsf{T}} = \bar{\mathbf{A}}$ and we show that then the numerator in (1) is real, which makes λ real. $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}$ is a scalar; hence taking the transpose has no effect. Thus

$$\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax} = (\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax})^{\mathsf{T}} = \mathbf{x}^{\mathsf{T}}\mathbf{A}^{\mathsf{T}}\bar{\mathbf{x}} = \mathbf{x}^{\mathsf{T}}\bar{\mathbf{A}}\bar{\mathbf{x}} = (\overline{\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}}). \tag{2}$$

Hence, $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}$ equals its complex conjugate, so that it must be real. ($a + ib = a - ib$ implies $b = 0$.)

(b) If $\mathbf{A}$ is skew-Hermitian, $\mathbf{A}^{\mathsf{T}} = -\bar{\mathbf{A}}$ and instead of (2) we obtain

$$\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax} = -(\overline{\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}}) \tag{3}$$

so that $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{Ax}$ equals minus its complex conjugate and is pure imaginary or 0. ($a + ib = -(a - ib)$ implies $a = 0$.)

(c) Let $\mathbf{A}$ be unitary. We take $\mathbf{Ax} = \lambda\mathbf{x}$ and its conjugate transpose

$$(\overline{\mathbf{Ax}})^{\mathsf{T}} = (\overline{\lambda\mathbf{x}})^{\mathsf{T}} = \bar{\lambda}\bar{\mathbf{x}}^{\mathsf{T}}$$

and multiply the two left sides and the two right sides,

$$(\overline{\mathbf{Ax}})^{\mathsf{T}}\mathbf{Ax} = \bar{\lambda}\lambda\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x} = |\lambda|^2\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}.$$

But $\mathbf{A}$ is unitary, $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$, so that on the left we obtain

$$(\overline{\mathbf{A}\mathbf{x}})^{\mathsf{T}}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}^{-1}\mathbf{A}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{I}\mathbf{x} = \bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}.$$

Together, $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x} = |\lambda|^2\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$. We now divide by $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}\ (\neq 0)$ to get $|\lambda|^2 = 1$. Hence $|\lambda| = 1$.
This proves Theorem 1 as well as Theorems 1 and 5 in Sec. 8.3. ■

Key properties of orthogonal matrices (invariance of the inner product, orthonormality of rows and columns; see Sec. 8.3) generalize to unitary matrices in a remarkable way.

To see this, instead of R^n we now use the **complex vector space** C^n of all complex vectors with n complex numbers as components, and complex numbers as scalars. For such complex vectors the **inner product** is defined by (note the overbar for the complex conjugate)

$$\mathbf{a} \cdot \mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b}. \tag{4}$$

The **length** or **norm** of such a complex vector is a *real* number defined by

$$\|\mathbf{a}\| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{\bar{\mathbf{a}}_j^{\mathsf{T}}\mathbf{a}} = \sqrt{\bar{a}_1 a_1 + \cdots + \bar{a}_n a_n} = \sqrt{|a_1|^2 + \cdots + |a_n|^2}. \tag{5}$$

THEOREM 2

Invariance of Inner Product

A **unitary transformation**, *that is,* $\mathbf{y} = \mathbf{A}\mathbf{x}$ *with a unitary matrix* $\mathbf{A}$, *preserves the value of the inner product* (4), *hence also the norm* (5).

PROOF The proof is the same as that of Theorem 2 in Sec. 8.3, which the theorem generalizes. In the analog of (9), Sec. 8.3, we now have bars,

$$\mathbf{u} \cdot \mathbf{v} = \bar{\mathbf{u}}^{\mathsf{T}}\mathbf{v} = (\overline{\mathbf{A}\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A}\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{I}\mathbf{b} = \bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = \mathbf{a} \cdot \mathbf{b}.$$

The complex analog of an orthonormal system of real vectors (see Sec. 8.3) is defined as follows.

DEFINITION

Unitary System

A *unitary system* is a set of complex vectors satisfying the relationships

$$\mathbf{a}_j \cdot \mathbf{a}_k = \bar{\mathbf{a}}_j^{\mathsf{T}}\mathbf{a}_k = \begin{cases} 0 & \text{if} \quad j \neq k \\ 1 & \text{if} \quad j = k. \end{cases} \tag{6}$$

Theorem 3 in Sec. 8.3 extends to complex as follows.

THEOREM 3

Unitary Systems of Column and Row Vectors

A complex square matrix is unitary if and only if its column vectors (and also its row vectors) form a unitary system.

PROOF The proof is the same as that of Theorem 3 in Sec. 8.3, except for the bars required in $\bar{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$ and in (4) and (6) of the present section.

THEOREM 4

Determinant of a Unitary Matrix

Let $\mathbf{A}$ *be a unitary matrix. Then its determinant has absolute value one, that is,* $|\det \mathbf{A}| = 1$.

PROOF Similarly, as in Sec. 8.3, we obtain

$$1 = \det(\mathbf{A}\mathbf{A}^{-1}) = \det(\mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}}) = \det \mathbf{A} \det \bar{\mathbf{A}}^{\mathsf{T}} = \det \mathbf{A} \det \bar{\mathbf{A}}$$
$$= \det \mathbf{A}\, \overline{\det \mathbf{A}} = |\det \mathbf{A}|^2.$$

Hence $|\det \mathbf{A}| = 1$ (where $\det \mathbf{A}$ may now be complex). ■

EXAMPLE 4 **Unitary Matrix Illustrating Theorems 1c and 2–4**

For the vectors $\mathbf{a}^{\mathsf{T}} = [2 \quad -i]$ and $\mathbf{b}^{\mathsf{T}} = [1 + i \quad 4i]$ we get $\bar{\mathbf{a}}^{\mathsf{T}} = [2 \quad i]^{\mathsf{T}}$ and $\bar{\mathbf{a}}^{\mathsf{T}}\mathbf{b} = 2(1 + i) - 4 = -2 + 2i$ and with

$$\mathbf{A} = \begin{bmatrix} 0.8i & 0.6 \\ 0.6 & 0.8i \end{bmatrix} \quad \text{also} \quad \mathbf{A}\mathbf{a} = \begin{bmatrix} i \\ 2 \end{bmatrix} \quad \text{and} \quad \mathbf{A}\mathbf{b} = \begin{bmatrix} -0.8 + 3.2i \\ -2.6 + 0.6i \end{bmatrix},$$

as one can readily verify. This gives $(\overline{\mathbf{A}\mathbf{a}})^{\mathsf{T}}\mathbf{A}\mathbf{b} = -2 + 2i$, illustrating Theorem 2. The matrix is unitary. Its columns form a unitary system,

$$\bar{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_1 = -0.8i \cdot 0.8i + 0.6^2 = 1, \qquad \bar{\mathbf{a}}_1^{\mathsf{T}}\mathbf{a}_2 = -0.8i \cdot 0.6 + 0.6 \cdot 0.8i = 0,$$
$$\bar{\mathbf{a}}_2^{\mathsf{T}}\mathbf{a}_2 = 0.6^2 + (-0.8i)0.8i = 1$$

and so do its rows. Also, $\det \mathbf{A} = -1$. The eigenvalues are $0.6 + 0.8i$ and $-0.6 + 0.8i$, with eigenvectors $[1 \quad 1]^{\mathsf{T}}$ and $[1 \quad -1]^{\mathsf{T}}$, respectively. ■

Theorem 2 in Sec. 8.4 on the existence of an eigenbasis extends to complex matrices as follows.

THEOREM 5

Basis of Eigenvectors

A Hermitian, skew-Hermitian, or unitary matrix has a basis of eigenvectors for C^n *that is a unitary system.*

For a proof see Ref. [B3], vol. 1, pp. 270–272 and p. 244 (Definition 2).

EXAMPLE 5 **Unitary Eigenbases**

The matrices $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in Example 2 have the following unitary systems of eigenvectors, as you should verify.

$$\mathbf{A}: \quad \frac{1}{\sqrt{35}}[1 - 3i \quad 5]^{\mathsf{T}} \quad (\lambda = 9), \qquad \frac{1}{\sqrt{14}}[1 - 3i \quad -2]^{\mathsf{T}} \quad (\lambda = 2)$$

$$\mathbf{B}: \quad \frac{1}{\sqrt{30}}[1 - 2i \quad -5]^{\mathsf{T}} \quad (\lambda = -2i), \qquad \frac{1}{\sqrt{30}}[5 \quad 1 + 2i]^{\mathsf{T}} \quad (\lambda = 4i)$$

$$\mathbf{C}: \quad \frac{1}{\sqrt{2}}[1 \quad 1]^{\mathsf{T}} \quad (\lambda = \tfrac{1}{2}(i + \sqrt{3})), \qquad \frac{1}{\sqrt{2}}[1 \quad -1]^{\mathsf{T}} \quad (\lambda = \tfrac{1}{2}(i - \sqrt{3})).$$

■

Hermitian and Skew-Hermitian Forms

The concept of a quadratic form (Sec. 8.4) can be extended to complex. We call the numerator $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x}$ in (1) a **form** in the components $x_1, \cdots, x_n$ of $\mathbf{x}$, which may now be complex. This form is again a sum of n^2 terms

$$\begin{aligned}\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x} &= \sum_{j=1}^{n}\sum_{k=1}^{n} a_{jk}\bar{x}_j x_k \\ &= \quad a_{11}\bar{x}_1x_1 + \cdots + a_{1n}\bar{x}_1x_n \\ &\quad + a_{21}\bar{x}_2x_1 + \cdots + a_{2n}\bar{x}_2x_n \\ &\quad + \cdots\cdots\cdots\cdots\cdots\cdots \\ &\quad + a_{n1}\bar{x}_nx_1 + \cdots + a_{nn}\bar{x}_nx_n.\end{aligned} \tag{7}$$

$\mathbf{A}$ is called its **coefficient matrix**. The form is called a **Hermitian** or **skew-Hermitian form** if $\mathbf{A}$ is Hermitian or skew-Hermitian, respectively. *The value of a Hermitian form is real, and that of a skew-Hermitian form is pure imaginary or zero.* This can be seen directly from (2) and (3) and accounts for the importance of these forms in physics. Note that (2) and (3) are valid for any vectors because, in the proof of (2) and (3), we did not use that $\mathbf{x}$ is an eigenvector but only that $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{x}$ is real and not 0.

EXAMPLE 6 Hermitian Form

For $\mathbf{A}$ in Example 2 and, say, $\mathbf{x} = [1+i \quad 5i]^{\mathsf{T}}$ we get

$$\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x} = [1-i \quad -5i]\begin{bmatrix} 4 & 1-3i \\ 1+3i & 7 \end{bmatrix}\begin{bmatrix} 1+i \\ 5i \end{bmatrix} = [1-i \quad -5i]\begin{bmatrix} 4(1+i)+(1-3i)\cdot 5i \\ (1+3i)(1+i)+7\cdot 5i \end{bmatrix} = 223. \qquad \blacksquare$$

Clearly, if $\mathbf{A}$ and $\mathbf{x}$ in (4) are real, then (7) reduces to a quadratic form, as discussed in the last section.

PROBLEM SET 8.5

1–6 EIGENVALUES AND VECTORS

Is the given matrix Hermitian? Skew-Hermitian? Unitary? Find its eigenvalues and eigenvectors.

1. $\begin{bmatrix} 6 & i \\ -i & 6 \end{bmatrix}$

2. $\begin{bmatrix} i & 1+i \\ -1+i & 0 \end{bmatrix}$

3. $\begin{bmatrix} \frac{1}{2} & i\sqrt{\frac{3}{4}} \\ i\sqrt{\frac{3}{4}} & \frac{1}{2} \end{bmatrix}$

4. $\begin{bmatrix} 0 & i \\ i & 0 \end{bmatrix}$

5. $\begin{bmatrix} i & 0 & 0 \\ 0 & 0 & i \\ 0 & i & 0 \end{bmatrix}$

6. $\begin{bmatrix} 0 & 2+2i & 0 \\ 2-2i & 0 & 2+2i \\ 0 & 2-2i & 0 \end{bmatrix}$

7. Pauli spin matrices. Find the eigenvalues and eigenvectors of the so-called *Pauli spin matrices* and show that $\mathbf{S}_x\mathbf{S}_y = i\mathbf{S}_z$, $\mathbf{S}_y\mathbf{S}_x = -i\mathbf{S}_z$, $\mathbf{S}_x^2 = \mathbf{S}_y^2 = \mathbf{S}_z^2 = \mathbf{I}$, where

$$\mathbf{S}_x = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \qquad \mathbf{S}_y = \begin{bmatrix} 0 & -i \\ i & 0 \end{bmatrix},$$

$$\mathbf{S}_z = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}.$$

8. Eigenvectors. Find eigenvectors of $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$ in Examples 2 and 3.

9–12 COMPLEX FORMS

Is the matrix **A** Hermitian or skew-Hermitian? Find $\bar{\mathbf{x}}^{\mathsf{T}}\mathbf{A}\mathbf{x}$. Show the details.

9. $\mathbf{A} = \begin{bmatrix} 4 & 3-2i \\ 3+2i & -4 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} -4i \\ 2+2i \end{bmatrix}$

10. $\mathbf{A} = \begin{bmatrix} i & -2+3i \\ 2+3i & 0 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 2i \\ 8 \end{bmatrix}$

11. $\mathbf{A} = \begin{bmatrix} i & 1 & 2+i \\ -1 & 0 & 3i \\ -2+i & 3i & i \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ i \\ -i \end{bmatrix}$

12. $\mathbf{A} = \begin{bmatrix} 1 & i & 4 \\ -i & 3 & 0 \\ 4 & 0 & 2 \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} 1 \\ i \\ -i \end{bmatrix}$

13–20 GENERAL PROBLEMS

13. Product. Show that $(\overline{\mathbf{ABC}})^{\mathsf{T}} = -\mathbf{C}^{-1}\mathbf{BA}$ for any $n \times n$ Hermitian **A**, skew-Hermitian **B**, and unitary **C**.

14. Product. Show $(\overline{\mathbf{BA}})^{\mathsf{T}} = -\mathbf{AB}$ for **A** and **B** in Example 2. For any $n \times n$ Hermitian **A** and skew-Hermitian **B**.

15. Decomposition. Show that any square matrix may be written as the sum of a Hermitian and a skew-Hermitian matrix. Give examples.

16. Unitary matrices. Prove that the product of two unitary $n \times n$ matrices and the inverse of a unitary matrix are unitary. Give examples.

17. Powers of unitary matrices in applications may sometimes be very simple. Show that $\mathbf{C}^{12} = \mathbf{I}$ in Example 2. Find further examples.

18. Normal matrix. This important concept denotes a matrix that commutes with its conjugate transpose, $\mathbf{A}\bar{\mathbf{A}}^{\mathsf{T}} = \bar{\mathbf{A}}^{\mathsf{T}}\mathbf{A}$. Prove that Hermitian, skew-Hermitian, and unitary matrices are normal. Give corresponding examples of your own.

19. Normality criterion. Prove that **A** is normal if and only if the Hermitian and skew-Hermitian matrices in Prob. 18 commute.

20. Find a simple matrix that is not normal. Find a normal matrix that is not Hermitian, skew-Hermitian, or unitary.

CHAPTER 8 REVIEW QUESTIONS AND PROBLEMS

1. In solving an eigenvalue problem, what is given and what is sought?

2. Give a few typical applications of eigenvalue problems.

3. Do there exist square matrices without eigenvalues?

4. Can a real matrix have complex eigenvalues? Can a complex matrix have real eigenvalues?

5. Does a 5×5 matrix always have a real eigenvalue?

6. What is algebraic multiplicity of an eigenvalue? Defect?

7. What is an eigenbasis? When does it exist? Why is it important?

8. When can we expect orthogonal eigenvectors?

9. State the definitions and main properties of the three classes of real matrices and of complex matrices that we have discussed.

10. What is diagonalization? Transformation to principal axes?

11–15 SPECTRUM

Find the eigenvalues. Find the eigenvectors.

11. $\begin{bmatrix} 2.5 & 0.5 \\ 0.5 & 2.5 \end{bmatrix}$

12. $\begin{bmatrix} -7 & 4 \\ -12 & 7 \end{bmatrix}$

13. $\begin{bmatrix} 8 & -1 \\ 5 & 2 \end{bmatrix}$

14. $\begin{bmatrix} 7 & 2 & -1 \\ 2 & 7 & 1 \\ -1 & 1 & 8.5 \end{bmatrix}$

15. $\begin{bmatrix} 0 & -3 & -6 \\ 3 & 0 & -6 \\ 6 & 6 & 0 \end{bmatrix}$

16–17 SIMILARITY

Verify that **A** and $\hat{\mathbf{A}} = \mathbf{p}^{-1}\mathbf{AP}$ have the same spectrum.

16. $\mathbf{A} = \begin{bmatrix} 19 & 12 \\ 12 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 2 & 4 \\ 4 & 2 \end{bmatrix}$

17. $\mathbf{A} = \begin{bmatrix} 7 & -4 \\ 12 & -7 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 5 & 3 \\ 3 & 5 \end{bmatrix}$

18. $\mathbf{A} = \begin{bmatrix} -4 & 6 & 6 \\ 0 & 2 & 0 \\ -1 & 1 & 1 \end{bmatrix}, \quad \mathbf{P} = \begin{bmatrix} 1 & 8 & -7 \\ 0 & 1 & 3 \\ 0 & 0 & 1 \end{bmatrix}$

19–21 DIAGONALIZATION

Find an eigenbasis and diagonalize.

9. $\begin{bmatrix} -1.4 & 1.0 \\ -1.0 & 1.1 \end{bmatrix}$

20. $\begin{bmatrix} 72 & -56 \\ -56 & 513 \end{bmatrix}$

21. $\begin{bmatrix} -12 & 22 & 6 \\ 8 & 2 & 6 \\ -8 & 20 & 16 \end{bmatrix}$

22–25 CONIC SECTIONS. PRINCIPAL AXES

Transform to canonical form (to principal axes). Express $[x_1 \quad x_2]^\mathsf{T}$ in terms of the new variables $[y_1 \quad y_2]^\mathsf{T}$.

22. $9x_1^2 - 6x_1x_2 + 17x_2^2 = 36$

23. $4x_1^2 + 24x_1x_2 - 14x_2^2 = 20$

24. $5x_1^2 + 24x_1x_2 - 5x_2^2 = 0$

25. $3.7x_1^2 + 3.2x_1x_2 + 1.3x_2^2 = 4.5$

SUMMARY OF CHAPTER 8

Linear Algebra: Matrix Eigenvalue Problems

The practical importance of matrix eigenvalue problems can hardly be overrated. The problems are defined by the vector equation

$$\mathbf{Ax} = \lambda \mathbf{x}. \tag{1}$$

A is a given square matrix. All matrices in this chapter are ***square***. λ is a scalar. To *solve* the problem (1) means to determine values of λ, called **eigenvalues** (or **characteristic values**) of **A**, such that (1) has a nontrivial solution **x** (that is, $\mathbf{x} \neq \mathbf{0}$), called an **eigenvector** of **A** corresponding to that λ. An $n \times n$ matrix has at least one and at most n numerically different eigenvalues. These are the solutions of the **characteristic equation** (Sec. 8.1)

$$D(\lambda) = \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0. \tag{2}$$

$D(\lambda)$ is called the **characteristic determinant** of **A**. By expanding it we get the **characteristic polynomial** of **A**, which is of degree n in λ. Some typical applications are shown in Sec. 8.2.

Section 8.3 is devoted to eigenvalue problems for **symmetric** ($\mathbf{A}^\mathsf{T} = \mathbf{A}$), **skew-symmetric** ($\mathbf{A}^\mathsf{T} = -\mathbf{A}$), and **orthogonal matrices** ($\mathbf{A}^\mathsf{T} = \mathbf{A}^{-1}$). Section 8.4 concerns the diagonalization of matrices and the transformation of quadratic forms to principal axes and its relation to eigenvalues.

Section 8.5 extends Sec. 8.3 to the complex analogs of those real matrices, called **Hermitian** ($\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}$), **skew-Hermitian** ($\bar{\mathbf{A}}^\mathsf{T} = -\mathbf{A}$), and **unitary matrices** ($\bar{\mathbf{A}}^\mathsf{T} = \mathbf{A}^{-1}$). All the eigenvalues of a Hermitian matrix (and a symmetric one) are real. For a skew-Hermitian (and a skew-symmetric) matrix they are pure imaginary or zero. For a unitary (and an orthogonal) matrix they have absolute value 1.

CHAPTER 9

Vector Differential Calculus. Grad, Div, Curl

Engineering, physics, and computer sciences, in general, but particularly solid mechanics, aerodynamics, aeronautics, fluid flow, heat flow, electrostatics, quantum physics, laser technology, robotics as well as other areas have applications that require an understanding of **vector calculus**. This field encompasses vector differential calculus and vector integral calculus. Indeed, the engineer, physicist, and mathematician need a good grounding in these areas as provided by the carefully chosen material of Chaps. 9 and 10.

Forces, velocities, and various other quantities may be thought of as vectors. Vectors appear frequently in the applications above and also in the biological and social sciences, so it is natural that problems are modeled in **3-space**. This is the space of three dimensions with the usual measurement of distance, as given by the Pythagorean theorem. Within that realm, **2-space** (the plane) is a special case. Working in 3-space requires that we extend the common differential calculus to vector differential calculus, that is, the calculus that deals with vector functions and vector fields and is explained in this chapter.

Chapter 9 is arranged in three groups of sections. Sections 9.1–9.3 extend the basic algebraic operations of vectors into 3-space. These operations include the inner product and the cross product. Sections 9.4 and 9.5 form the heart of vector differential calculus. Finally, Secs. 9.7–9.9 discuss three physically important concepts related to scalar and vector fields: gradient (Sec. 9.7), divergence (Sec. 9.8), and curl (Sec. 9.9). They are expressed in Cartesian coordinates in this chapter and, if desired, expressed in ***curvilinear coordinates*** in a short section in App. A3.4.

We shall keep this chapter ***independent of Chaps. 7 and 8***. Our present approach is in harmony with Chap. 7, with the restriction to two and three dimensions providing for a richer theory with basic physical, engineering, and geometric applications.

Prerequisite: Elementary use of second- and third-order determinants in Sec. 9.3.
Sections that may be omitted in a shorter course: 9.5, 9.6.
References and Answers to Problems: App. 1 Part B, App. 2.

9.1 Vectors in 2-Space and 3-Space

In engineering, physics, mathematics, and other areas we encounter two kinds of quantities. They are scalars and vectors.

A **scalar** is a quantity that is determined by its magnitude. It takes on a numerical value, i.e., a number. Examples of scalars are time, temperature, length, distance, speed, density, energy, and voltage.

In contrast, a **vector** is a quantity that has both magnitude and direction. We can say that a vector is an ***arrow*** or a ***directed line segment***. For example, a velocity vector has length or magnitude, which is speed, and direction, which indicates the direction of motion. Typical examples of vectors are displacement, velocity, and force, see Fig. 164 as an illustration.

More formally, we have the following. We denote vectors by lowercase boldface letters **a**, **b**, **v**, etc. In handwriting you may use arrows, for instance, $\vec{a}$ (in place of **a**), $\vec{b}$, etc.

A vector (arrow) has a tail, called its **initial point**, and a tip, called its **terminal point**. This is motivated in the **translation** (displacement without rotation) of the triangle in Fig. 165, where the initial point P of the vector **a** is the original position of a point, and the terminal point Q is the terminal position of that point, its position *after* the translation. The length of the arrow equals the distance between P and Q. This is called the **length** (or *magnitude*) of the vector **a** and is denoted by $|\mathbf{a}|$. Another name for *length* is **norm** (or *Euclidean norm*).

A vector of length 1 is called a **unit vector**.

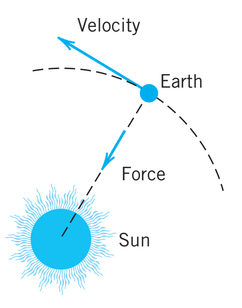


Fig. 164. Force and velocity

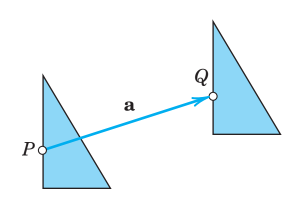


Fig. 165. Translation

Of course, we would like to calculate with vectors. For instance, we want to find the resultant of forces or compare parallel forces of different magnitude. This motivates our next ideas: to define *components* of a vector, and then the two basic algebraic operations of *vector addition* and *scalar multiplication.*

For this we must first define *equality of vectors* in a way that is practical in connection with forces and other applications.

DEFINITION

Equality of Vectors

Two vectors **a** and **b** are equal, written $\mathbf{a} = \mathbf{b}$, if they have the same length and the same direction [as explained in Fig. 166; in particular, note (B)]. Hence a vector can be arbitrarily translated; that is, its initial point can be chosen arbitrarily.

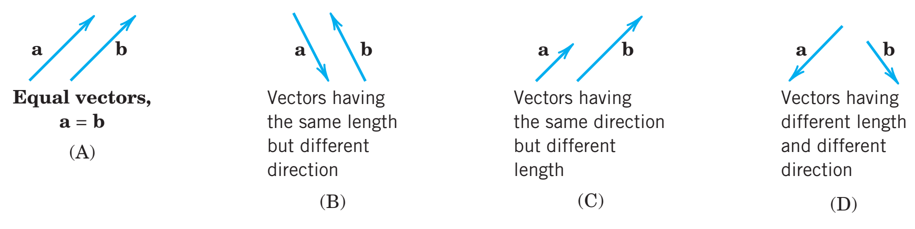


Fig. 166. (A) Equal vectors. (B)–(D) Different vectors

Components of a Vector

We choose an xyz **Cartesian coordinate system**[1] in space (Fig. 167), that is, a usual rectangular coordinate system with the same scale of measurement on the three mutually perpendicular coordinate axes. Let **a** be a given vector with initial point P: (x_1, y_1, z_1) and terminal point Q: (x_2, y_2, z_2). Then the three coordinate differences

(1) $$a_1 = x_2 - x_1, \qquad a_2 = y_2 - y_1, \qquad a_3 = z_2 - z_1$$

are called the **components** of the vector **a** with respect to that coordinate system, and we write simply $\mathbf{a} = [a_1, a_2, a_3]$. See Fig. 168.

The **length** $|\mathbf{a}|$ of **a** can now readily be expressed in terms of components because from (1) and the Pythagorean theorem we have

(2) $$|\mathbf{a}| = \sqrt{a_1^2 + a_2^2 + a_3^2}.$$

EXAMPLE 1 Components and Length of a Vector

The vector **a** with initial point P: (4, 0, 2) and terminal point Q: (6, −1, 2) has the components

$$a_1 = 6 - 4 = 2, \qquad a_2 = -1 - 0 = -1, \qquad a_3 = 2 - 2 = 0.$$

Hence $\mathbf{a} = [2, -1, 0]$. (Can you sketch **a**, as in Fig. 168?) Equation (2) gives the length

$$|\mathbf{a}| = \sqrt{2^2 + (-1)^2 + 0^2} = \sqrt{5}.$$

If we choose (−1, 5, 8) as the initial point of **a**, the corresponding terminal point is (1, 4, 8).

If we choose the origin (0, 0, 0) as the initial point of **a**, the corresponding terminal point is (2, −1, 0); its coordinates equal the components of **a**. This suggests that we can determine each point in space by a vector, called the *position vector* of the point, as follows. ■

A Cartesian coordinate system being given, the **position vector r** of a point A: (x, y, z) is the vector with the origin (0, 0, 0) as the initial point and A as the terminal point (see Fig. 169). Thus in components, $\mathbf{r} = [x, y, z]$. This can be seen directly from (1) with $x_1 = y_1 = z_1 = 0$.

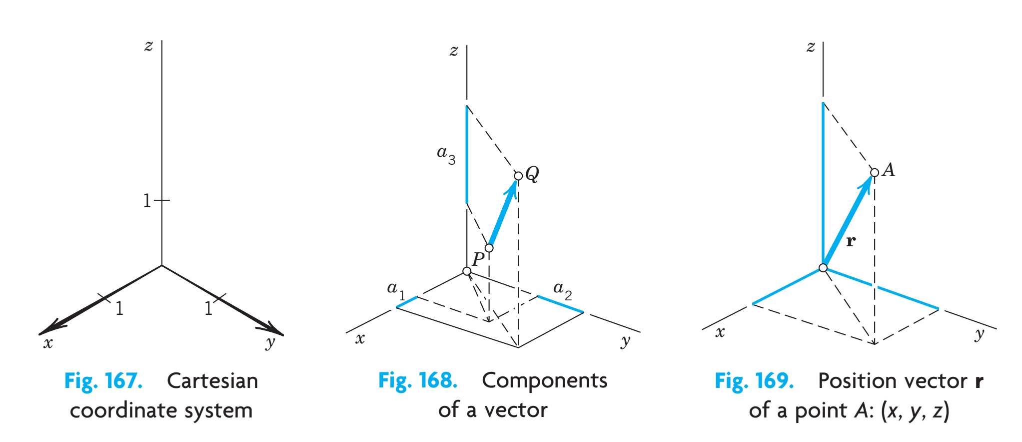


Fig. 167. Cartesian coordinate system

Fig. 168. Components of a vector

Fig. 169. Position vector **r** of a point A: (x, y, z)

[1]Named after the French philosopher and mathematician RENATUS CARTESIUS, latinized for RENÉ DESCARTES (1596–1650), who invented analytic geometry. His basic work *Géométrie* appeared in 1637, as an appendix to his *Discours de la méthode*.

Furthermore, if we translate a vector **a**, with initial point P and terminal point Q, then corresponding coordinates of P and Q change by the same amount, so that the differences in (1) remain unchanged. This proves

THEOREM 1

Vectors as Ordered Triples of Real Numbers

A fixed Cartesian coordinate system being given, each vector is uniquely determined by its ordered triple of corresponding components. Conversely, to each ordered triple of real numbers (a_1, a_2, a_3) *there corresponds precisely one vector* $\mathbf{a} = [a_1, a_2, a_3]$, *with* $(0, 0, 0)$ *corresponding to the* **zero vector 0**, *which has length* 0 *and no direction.*

Hence a vector equation $\mathbf{a} = \mathbf{b}$ *is equivalent to the three equations* $a_1 = b_1$, $a_2 = b_2$, $a_3 = b_3$ *for the components.*

We now see that from our "geometric" definition of a vector as an arrow we have arrived at an "algebraic" characterization of a vector by Theorem 1. We could have started from the latter and reversed our process. This shows that the two approaches are equivalent.

Vector Addition, Scalar Multiplication

Calculations with vectors are very useful and are almost as simple as the arithmetic for real numbers. Vector arithmetic follows almost naturally from applications. We first define how to add vectors and later on how to multiply a vector by a number.

DEFINITION

Addition of Vectors

The **sum a + b** of two vectors $\mathbf{a} = [a_1, a_2, a_3]$ and $\mathbf{b} = [b_1, b_2, b_3]$ is obtained by adding the corresponding components,

$$\mathbf{a} + \mathbf{b} = [a_1 + b_1, \quad a_2 + b_2, \quad a_3 + b_3]. \tag{3}$$

Geometrically, place the vectors as in Fig. 170 (the initial point of **b** at the terminal point of **a**); then **a + b** is the vector drawn from the initial point of **a** to the terminal point of **b**.

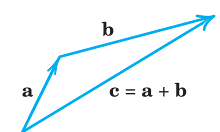


Fig. 170. Vector addition

For forces, this addition is the parallelogram law by which we obtain the **resultant** of two forces in mechanics. See Fig. 171.

Figure 172 shows (for the plane) that the "algebraic" way and the "geometric way" of vector addition give the same vector.

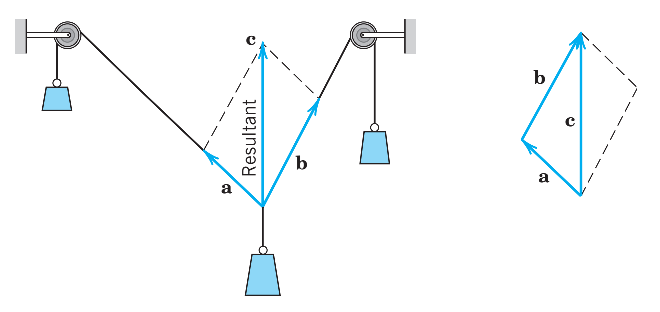


Fig. 171. Resultant of two forces (**parallelogram law**)

Basic Properties of Vector Addition. Familiar laws for real numbers give immediately

$$\text{(4)}\quad \begin{aligned} &\text{(a)} \quad \mathbf{a} + \mathbf{b} = \mathbf{b} + \mathbf{a} && \textit{(Commutativity)} \\ &\text{(b)} \quad (\mathbf{u} + \mathbf{v}) + \mathbf{w} = \mathbf{u} + (\mathbf{v} + \mathbf{w}) && \textit{(Associativity)} \\ &\text{(c)} \quad \mathbf{a} + \mathbf{0} = \mathbf{0} + \mathbf{a} = \mathbf{a} \\ &\text{(d)} \quad \mathbf{a} + (-\mathbf{a}) = \mathbf{0}. \end{aligned}$$

Properties (a) and (b) are verified geometrically in Figs. 173 and 174. Furthermore, $-\mathbf{a}$ denotes the vector having the length $|\mathbf{a}|$ and the direction opposite to that of $\mathbf{a}$.

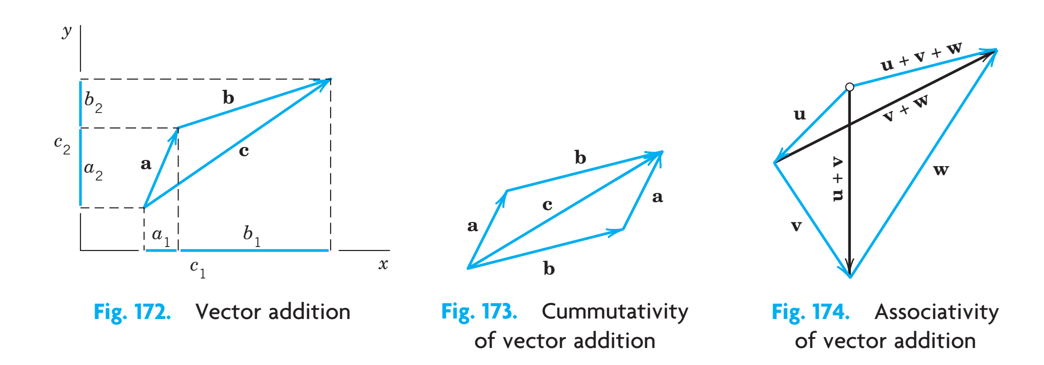


Fig. 172. Vector addition

Fig. 173. Cummutativity of vector addition

Fig. 174. Associativity of vector addition

In (4b) we may simply write $\mathbf{u} + \mathbf{v} + \mathbf{w}$, and similarly for sums of more than three vectors. Instead of $\mathbf{a} + \mathbf{a}$ we also write $2\mathbf{a}$, and so on. This (and the notation $-\mathbf{a}$ used just before) motivates defining the second algebraic operation for vectors as follows.

DEFINITION

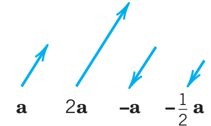


Fig. 175. Scalar multiplication [multiplication of vectors by scalars (numbers)]

Scalar Multiplication (Multiplication by a Number)

The product $c\mathbf{a}$ of any vector $\mathbf{a} = [a_1, a_2, a_3]$ and any scalar c (real number c) is the vector obtained by multiplying each component of $\mathbf{a}$ by c,

$$\text{(5)}\qquad c\mathbf{a} = [ca_1, ca_2, ca_3].$$

Geometrically, if $\mathbf{a} \neq \mathbf{0}$, then $c\mathbf{a}$ with $c > 0$ has the direction of $\mathbf{a}$ and with $c < 0$ the direction opposite to $\mathbf{a}$. In any case, the length of $c\mathbf{a}$ is $|c\mathbf{a}| = |c||\mathbf{a}|$, and $c\mathbf{a} = \mathbf{0}$ if $\mathbf{a} = \mathbf{0}$ or $c = 0$ (or both). (See Fig. 175.)

Basic Properties of Scalar Multiplication. From the definitions we obtain directly

$$\text{(6)}\quad \begin{aligned} &\text{(a)} \quad c(\mathbf{a} + \mathbf{b}) = c\mathbf{a} + c\mathbf{b} \\ &\text{(b)} \quad (c + k)\mathbf{a} = c\mathbf{a} + k\mathbf{a} \\ &\text{(c)} \quad c(k\mathbf{a}) = (ck)\mathbf{a} && \text{(written } ck\mathbf{a}\text{)} \\ &\text{(d)} \quad 1\mathbf{a} = \mathbf{a}. \end{aligned}$$

You may prove that (4) and (6) imply for any vector **a**

$$\text{(7)} \qquad \begin{aligned} &\text{(a)} \quad 0\mathbf{a} = \mathbf{0} \\ &\text{(b)} \quad (-1)\mathbf{a} = -\mathbf{a}. \end{aligned}$$

Instead of $\mathbf{b} + (-\mathbf{a})$ we simply write $\mathbf{b} - \mathbf{a}$ (Fig. 176).

EXAMPLE 2 Vector Addition. Multiplication by Scalars

With respect to a given coordinate system, let

$$\mathbf{a} = [4, 0, 1] \quad \text{and} \quad \mathbf{b} = [2, -5, \tfrac{1}{3}].$$

Then $-\mathbf{a} = [-4, 0, -1]$, $7\mathbf{a} = [28, 0, 7]$, $\mathbf{a} + \mathbf{b} = [6, -5, \tfrac{4}{3}]$, and

$$2(\mathbf{a} - \mathbf{b}) = 2[2, 5, \tfrac{2}{3}] = [4, 10, \tfrac{4}{3}] = 2\mathbf{a} - 2\mathbf{b}.$$ ■

Unit Vectors i, j, k. Besides $\mathbf{a} = [a_1, a_2, a_3]$ another popular way of writing vectors is

$$\text{(8)} \qquad \mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}.$$

In this representation, **i**, **j**, **k** are the unit vectors in the positive directions of the axes of a Cartesian coordinate system (Fig. 177). Hence, in components,

$$\text{(9)} \qquad \mathbf{i} = [1, 0, 0], \quad \mathbf{j} = [0, 1, 0], \quad \mathbf{k} = [0, 0, 1]$$

and the right side of (8) is a sum of three vectors parallel to the three axes.

EXAMPLE 3 ijk Notation for Vectors

In Example 2 we have $\mathbf{a} = 4\mathbf{i} + \mathbf{k}$, $\mathbf{b} = 2\mathbf{i} - 5\mathbf{j} + \tfrac{1}{3}\mathbf{k}$, and so on. ■

All the vectors $\mathbf{a} = [a_1, a_2, a_3] = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ (with real numbers as components) form the **real vector space** R^3 with the two *algebraic operations* of vector addition and scalar multiplication as just defined. R^3 has **dimension** 3. The triple of vectors **i**, **j**, **k** is called a **standard basis** of R^3. Given a Cartesian coordinate system, the representation (8) of a given vector is unique.

Vector space R^3 is a model of a general vector space, as discussed in Sec. 7.9, but is not needed in this chapter.

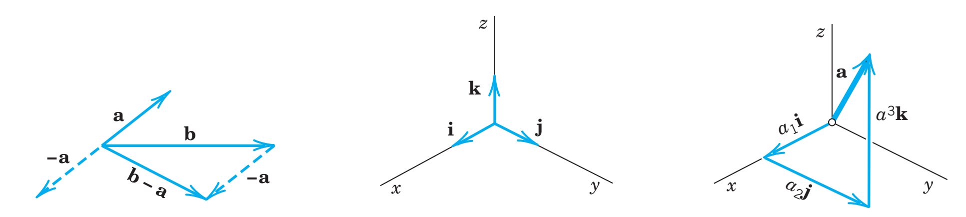


Fig. 176. Difference of vectors

Fig. 177. The unit vectors **i**, **j**, **k** and the representation (8)

PROBLEM SET 9.1

1–5 COMPONENTS AND LENGTH

Find the components of the vector **v** with initial point P and terminal point Q. Find $|\mathbf{v}|$. Sketch $|\mathbf{v}|$. Find the unit vector **u** in the direction of **v**.

1. P: (1, 1, 0), Q: (6, 2, 0)

2. P: (1, 1, 1), Q: (2, 2, 0)

3. P: (−3.0, 4,0, −0.5), Q: (5.5, 0, 1.2)

4. P: (1, 4, 2), Q: (−1, −4, −2)

5. P: (0, 0, 0), Q: (2, 1, −2)

6–10

Find the terminal point Q of the vector **v** with components as given and initial point P. Find $|\mathbf{v}|$.

6. 4, 0, 0; P: (0, 2, 13)

7. $\frac{1}{2}, 3, -\frac{1}{4}$; P: $(\frac{7}{2}, -3, \frac{3}{4})$

8. 13.1, 0.8, −2.0; P: (0, 0, 0)

9. 6, 1, −4; P: (−6, −1, −4)

10. 0, −3, 3; P: (0, 3, −3)

11–18 ADDITION, SCALAR MULTIPLICATION

Let $\mathbf{a} = [3, 2, 0] = 3\mathbf{i} + 2\mathbf{j}$; $\mathbf{b} = [-4, 6, 0] = 4\mathbf{i} + 6\mathbf{j}$, $\mathbf{c} = [5, -1, 8] = 5\mathbf{i} - \mathbf{j} + 8\mathbf{k}$, $\mathbf{d} = [0, 0, 4] = 4\mathbf{k}$. Find:

11. $2\mathbf{a}$, $\frac{1}{2}\mathbf{a}$, $-\mathbf{a}$

12. $(\mathbf{a} + \mathbf{b}) + \mathbf{c}$, $\mathbf{a} + (\mathbf{b} + \mathbf{c})$

13. $\mathbf{b} + \mathbf{c}$, $\mathbf{c} + \mathbf{b}$

14. $3\mathbf{c} - 6\mathbf{d}$, $3(\mathbf{c} - 2\mathbf{d})$

15. $7(\mathbf{c} - \mathbf{b})$, $7\mathbf{c} - 7\mathbf{b}$

16. $\frac{9}{2}\mathbf{a} - 3\mathbf{c}$, $9(\frac{1}{2}\mathbf{a} - \frac{1}{3}\mathbf{c})$

17. $(7 - 3)\mathbf{a}$, $7\mathbf{a} - 3\mathbf{a}$

18. $4\mathbf{a} + 3\mathbf{b}$, $-4\mathbf{a} - 3\mathbf{b}$

19. What laws do Probs. 12–16 illustrate?

20. Prove Eqs. (4) and (6).

21–25 FORCES, RESULTANT

Find the resultant in terms of components and its magnitude.

21. $\mathbf{p} = [2, 3, 0]$, $\mathbf{q} = [0, 6, 1]$, $\mathbf{u} = [2, 0, -4]$

22. $\mathbf{p} = [1, -2, 3]$, $\mathbf{q} = [3, 21, -16]$, $\mathbf{u} = [-4, -19, 13]$

23. $\mathbf{u} = [8, -1, 0]$, $\mathbf{v} = [\frac{1}{2}, 0, \frac{4}{3}]$, $\mathbf{w} = [-\frac{17}{2}, 1, \frac{11}{3}]$

24. $\mathbf{p} = [-1, 2, -3]$, $\mathbf{q} = [1, 1, 1]$, $\mathbf{u} = [1, -2, 2]$

25. $\mathbf{u} = [3, 1, -6]$, $\mathbf{v} = [0, 2, 5]$, $\mathbf{w} = [3, -1, -13]$

26–37 FORCES, VELOCITIES

26. Equilibrium. Find **v** such that **p**, **q**, **u** in Prob. 21 and **v** are in equilibrium.

27. Find **p** such that **u**, **v**, **w** in Prob. 23 and **p** are in equilibrium.

28. Unit vector. Find the unit vector in the direction of the resultant in Prob. 24.

29. Restricted resultant. Find all **v** such that the resultant of **v**, **p**, **q**, **u** with **p**, **q**, **u** as in Prob. 21 is parallel to the xy-plane.

30. Find **v** such that the resultant of **p**, **q**, **u**, **v** with **p**, **q**, **u** as in Prob. 24 has no components in x- and y-directions.

31. For what k is the resultant of [2, 0, −7], [1, 2, −3], and [0, 3, k] parallel to the xy-plane?

32. If $|\mathbf{p}| = 6$ and $|\mathbf{q}| = 4$, what can you say about the magnitude and direction of the resultant? Can you think of an application to robotics?

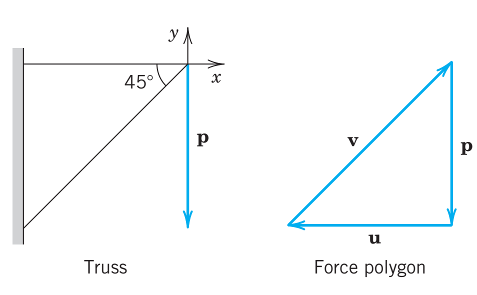
33. Same question as in Prob. 32 if $|\mathbf{p}| = 9$, $|\mathbf{q}| = 6$, $|\mathbf{u}| = 3$.

34. Relative velocity. If airplanes A and B are moving southwest with speed $|\mathbf{v}_A| = 550$ mph, and northwest with speed $|\mathbf{v}_B| = 450$ mph, respectively, what is the relative velocity $\mathbf{v} = \mathbf{v}_B - \mathbf{v}_A$ of B with respect to A?

35. Same question as in Prob. 34 for two ships moving northeast with speed $|\mathbf{v}_A| = 22$ knots and west with speed $|\mathbf{v}_B| = 19$ knots.

36. Reflection. If a ray of light is reflected once in each of two mutually perpendicular mirrors, what can you say about the reflected ray?

37. Force polygon. Truss. Find the forces in the system of two rods (*truss*) in the figure, where $|\mathbf{p}| = 1000$ nt. *Hint.* Forces in equilibrium form a polygon, the *force polygon*.



Problem 37

38. TEAM PROJECT. Geometric Applications. To increase your skill in dealing with vectors, use vectors to prove the following (see the figures).

(a) The diagonals of a parallelogram bisect each other.

(b) The line through the midpoints of adjacent sides of a parallelogram bisects one of the diagonals in the ratio 1 : 3.

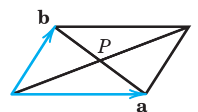
(c) Obtain (b) from (a).

(d) The three medians of a triangle (the segments from a vertex to the midpoint of the opposite side) meet at a single point, which divides the medians in the ratio 2 : 1.

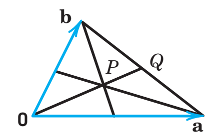
(e) The quadrilateral whose vertices are the midpoints of the sides of an arbitrary quadrilateral is a parallelogram.

(f) The four space diagonals of a parallelepiped meet and bisect each other.

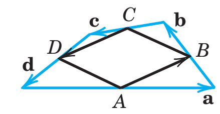
(g) The sum of the vectors drawn from the center of a regular polygon to its vertices is the zero vector.



Team Project 38(a)



Team Project 38(d)



Team Project 38(e)

9.2 Inner Product (Dot Product)

Orthogonality

The inner product or dot product can be motivated by calculating work done by a constant force, determining components of forces, or other applications. It involves the length of vectors and the angle between them. The inner product is a kind of multiplication of two vectors, defined in such a way that the outcome is a scalar. Indeed, another term for inner product is scalar product, a term we shall not use here. The definition of the inner product is as follows.

DEFINITION

Inner Product (Dot Product) of Vectors

The **inner product** or **dot product** $\mathbf{a} \bullet \mathbf{b}$ (read "**a** dot **b**") of two vectors **a** and **b** is the product of their lengths times the cosine of their angle (see Fig. 178),

$$(1) \qquad \begin{aligned} \mathbf{a} \bullet \mathbf{b} &= |\mathbf{a}||\mathbf{b}| \cos \gamma \qquad &\text{if} \quad \mathbf{a} \neq \mathbf{0}, \mathbf{b} \neq \mathbf{0} \\ \mathbf{a} \bullet \mathbf{b} &= 0 &\text{if} \quad \mathbf{a} = \mathbf{0} \text{ or } \mathbf{b} = \mathbf{0}. \end{aligned}$$

The angle γ, $0 \leqq \gamma \leqq \pi$, between **a** and **b** is measured when the initial points of the vectors coincide, as in Fig. 178. In components, $\mathbf{a} = [a_1, a_2, a_3]$, $\mathbf{b} = [b_1, b_2, b_3]$, and

$$(2) \qquad \mathbf{a} \bullet \mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3.$$

The second line in (1) is needed because γ is undefined when $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$. The derivation of (2) from (1) is shown below.

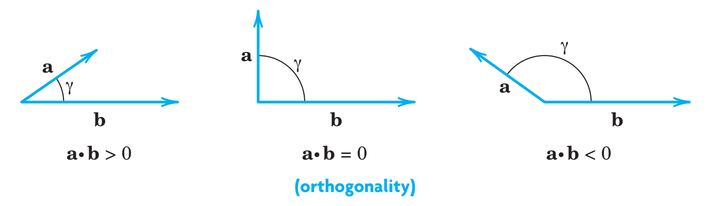


Fig. 178. Angle between vectors and value of inner product

Orthogonality. Since the cosine in (1) may be positive, 0, or negative, so may be the inner product (Fig. 178). The case that the inner product is zero is of particular practical interest and suggests the following concept.

A vector **a** is called **orthogonal** to a vector **b** if $\mathbf{a} \cdot \mathbf{b} = 0$. Then **b** is also orthogonal to **a**, and we call **a** and **b** **orthogonal vectors**. Clearly, this happens for nonzero vectors if and only if $\cos \gamma = 0$; thus $\gamma = \pi/2$ (90°). This proves the important

THEOREM 1

Orthogonality Criterion

The inner product of two nonzero vectors is 0 *if and only if these vectors are perpendicular.*

Length and Angle. Equation (1) with $\mathbf{b} = \mathbf{a}$ gives $\mathbf{a} \cdot \mathbf{a} = |\mathbf{a}|^2$. Hence

$$|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}}. \tag{3}$$

From (3) and (1) we obtain for the angle γ between two nonzero vectors

$$\cos \gamma = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \frac{\mathbf{a} \cdot \mathbf{b}}{\sqrt{\mathbf{a} \cdot \mathbf{a}}\sqrt{\mathbf{b} \cdot \mathbf{b}}}. \tag{4}$$

EXAMPLE 1 **Inner Product. Angle Between Vectors**

Find the inner product and the lengths of $\mathbf{a} = [1, 2, 0]$ and $\mathbf{b} = [3, -2, 1]$ as well as the angle between these vectors.

Solution. $\mathbf{a} \cdot \mathbf{b} = 1 \cdot 3 + 2 \cdot (-2) + 0 \cdot 1 = -1$, $|\mathbf{a}| = \sqrt{\mathbf{a} \cdot \mathbf{a}} = \sqrt{5}$, $|\mathbf{b}| = \sqrt{\mathbf{b} \cdot \mathbf{b}} = \sqrt{14}$, and (4) gives the angle

$$\gamma = \arccos \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} = \arccos(-0.11952) = 1.69061 = 96.865°.$$ ■

From the definition we see that the inner product has the following properties. For any vectors **a**, **b**, **c** and scalars q_1, q_2,

$$\textbf{(5)}\quad \begin{array}{ll} \text{(a)} & (q_1\mathbf{a} + q_2\mathbf{b}) \cdot \mathbf{c} = q_1\mathbf{a} \cdot \mathbf{c} + q_1\mathbf{b} \cdot \mathbf{c} \qquad (\textit{Linearity}) \\ \text{(b)} & \mathbf{a} \cdot \mathbf{b} = \mathbf{b} \cdot \mathbf{a} \qquad (\textit{Symmetry}) \\ \text{(c)} & \left.\begin{array}{l} \mathbf{a} \cdot \mathbf{a} \geqq 0 \\ \mathbf{a} \cdot \mathbf{a} = 0 \quad \text{if and only if} \quad \mathbf{a} = \mathbf{0} \end{array}\right\} \quad (\textit{Positive-definiteness}). \end{array}$$

Hence *dot multiplication is commutative* as shown by (5b). Furthermore, it is *distributive with respect to vector addition.* This follows from (5a) with $q_1 = 1$ and $q_2 = 1$:

$$\textbf{(5a*)}\qquad (\mathbf{a} + \mathbf{b}) \cdot \mathbf{c} = \mathbf{a} \cdot \mathbf{c} + \mathbf{b} \cdot \mathbf{c} \qquad (\textit{Distributivity}).$$

Furthermore, from (1) and $|\cos \gamma| \leqq 1$ we see that

$$\textbf{(6)}\qquad |\mathbf{a} \cdot \mathbf{b}| \leqq |\mathbf{a}||\mathbf{b}| \qquad (\textit{Cauchy–Schwarz inequality}).$$

Using this and (3), you may prove (see Prob. 16)

$$\textbf{(7)}\qquad |\mathbf{a} + \mathbf{b}| \leqq |\mathbf{a}| + |\mathbf{b}| \qquad (\textbf{Triangle inequality}).$$

Geometrically, (7) with $<$ says that one side of a triangle must be shorter than the other two sides together; this motivates the name of (7).

A simple direct calculation with inner products shows that

$$(8)\qquad |\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = 2(|\mathbf{a}|^2 + |\mathbf{b}|^2) \quad (\textit{Parallelogram equality}).$$

Equations (6)–(8) play a basic role in so-called *Hilbert spaces*, which are abstract inner product spaces. Hilbert spaces form the basis of quantum mechanics, for details see [GenRef7] listed in App. 1.

Derivation of (2) from (1). We write $\mathbf{a} = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ and $\mathbf{b} = b_1\mathbf{i} + b_2\mathbf{j} + b_3\mathbf{k}$, as in (8) of Sec. 9.1. If we substitute this into $\mathbf{a} \cdot \mathbf{b}$ and use (5a*), we first have a sum of $3 \times 3 = 9$ products

$$\mathbf{a} \cdot \mathbf{b} = a_1b_1\mathbf{i} \cdot \mathbf{i} + a_1b_2\mathbf{i} \cdot \mathbf{j} + \cdots + a_3b_3\mathbf{k} \cdot \mathbf{k}.$$

Now **i**, **j**, **k** are unit vectors, so that $\mathbf{i} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{k} = 1$ by (3). Since the coordinate axes are perpendicular, so are **i**, **j**, **k**, and Theorem 1 implies that the other six of those nine products are 0, namely, $\mathbf{i} \cdot \mathbf{j} = \mathbf{j} \cdot \mathbf{i} = \mathbf{j} \cdot \mathbf{k} = \mathbf{k} \cdot \mathbf{j} = \mathbf{k} \cdot \mathbf{i} = \mathbf{i} \cdot \mathbf{k} = 0$. But this reduces our sum for $\mathbf{a} \cdot \mathbf{b}$ to (2). ■

Applications of Inner Products

Typical applications of inner products are shown in the following examples and in Problem Set 9.2.

EXAMPLE 2 **Work Done by a Force Expressed as an Inner Product**

This is a major application. It concerns a body on which a *constant* force **p** acts. (For a *variable* force, see Sec. 10.1.) Let the body be given a displacement **d**. Then the work done by **p** in the displacement is defined as

$$W = |\mathbf{p}||\mathbf{d}| \cos \alpha = \mathbf{p} \cdot \mathbf{d}, \tag{9}$$

that is, magnitude $|\mathbf{p}|$ of the force times length $|\mathbf{d}|$ of the displacement times the cosine of the angle α between **p** and **d** (Fig. 179). If $\alpha < 90°$, as in Fig. 179, then $W > 0$. If **p** and **d** are orthogonal, then the work is zero (why?). If $\alpha > 90°$, then $W < 0$, which means that in the displacement one has to do work against the force. For example, think of swimming across a river at some angle α against the current.

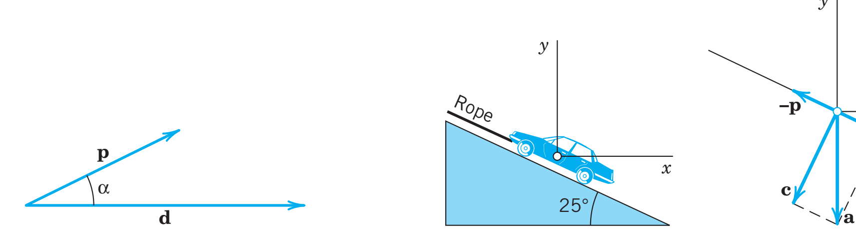


Fig. 179. Work done by a force

Fig. 180. Example 3

EXAMPLE 3 **Component of a Force in a Given Direction**

What force in the rope in Fig. 180 will hold a car of 5000 lb in equilibrium if the ramp makes an angle of 25° with the horizontal?

Solution. Introducing coordinates as shown, the weight is $\mathbf{a} = [0, -5000]$ because this force points downward, in the negative y-direction. We have to represent **a** as a sum (resultant) of two forces, $\mathbf{a} = \mathbf{c} + \mathbf{p}$, where **c** is the force the car exerts on the ramp, which is of no interest to us, and **p** is parallel to the rope. A vector in the direction of the rope is (see Fig. 180)

$$\mathbf{b} = [-1, \tan 25°] = [-1, 0.46631], \qquad \text{thus} \qquad |\mathbf{b}| = 1.10338,$$

The direction of the unit vector **u** is opposite to the direction of the rope so that

$$\mathbf{u} = -\frac{1}{|\mathbf{b}|}\mathbf{b} = [0.90631, -0.42262].$$

Since $|\mathbf{u}| = 1$ and $\cos \gamma > 0$, we see that we can write our result as

$$|\mathbf{p}| = (|\mathbf{a}| \cos \gamma)|\mathbf{u}| = \mathbf{a} \cdot \mathbf{u} = -\frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|} = \frac{5000 \cdot 0.46631}{1.10338} = 2113 \text{ [lb]}.$$

We can also note that $\gamma = 90° - 25° = 65°$ is the angle between **a** and **p** so that

$$|\mathbf{p}| = |\mathbf{a}| \cos \gamma = 5000 \cos 65° = 2113 \text{ [lb]}.$$

Answer: About 2100 lb. ■

Example 3 is typical of applications that deal with the **component** or **projection** *of a vector* **a** *in the direction of a vector* **b** $(\neq \mathbf{0})$. If we denote by p the length of the orthogonal projection of **a** on a straight line l parallel to **b** as shown in Fig. 181, then

$$p = |\mathbf{a}| \cos \gamma. \tag{10}$$

Here p is taken with the plus sign if $p\mathbf{b}$ has the direction of **b** and with the minus sign if $p\mathbf{b}$ has the direction opposite to **b**.

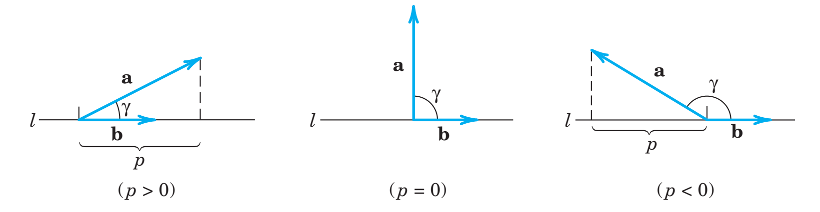


Fig. 181. Component of a vector **a** in the direction of a vector **b**

Multiplying (10) by $|\mathbf{b}|/|\mathbf{b}| = 1$, we have $\mathbf{a} \cdot \mathbf{b}$ in the numerator and thus

$$p = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{b}|} \qquad (\mathbf{b} \neq \mathbf{0}). \tag{11}$$

If **b** is a unit vector, as it is often used for fixing a direction, then (11) simply gives

$$p = \mathbf{a} \cdot \mathbf{b} \qquad (|\mathbf{b}| = 1). \tag{12}$$

Figure 182 shows the projection p of **a** in the direction of **b** (as in Fig. 181) and the projection $q = |\mathbf{b}| \cos \gamma$ of **b** in the direction of **a**.

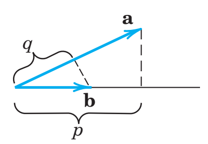


Fig. 182. Projections p of **a** on **b** and q of **b** on **a**

EXAMPLE 4 Orthonormal Basis

By definition, an *orthonormal basis* for 3-space is a basis $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ consisting of orthogonal unit vectors. It has the great advantage that the determination of the coefficients in representations $\mathbf{v} = l_1\mathbf{a} + l_2\mathbf{b} + l_3\mathbf{c}$ of a given vector **v** is very simple. We claim that $l_1 = \mathbf{a} \cdot \mathbf{v}$, $l_2 = \mathbf{b} \cdot \mathbf{v}$, $l_3 = \mathbf{c} \cdot \mathbf{v}$. Indeed, this follows simply by taking the inner products of the representation with **a**, **b**, **c**, respectively, and using the orthonormality of the basis, $\mathbf{a} \cdot \mathbf{v} = l_1\mathbf{a} \cdot \mathbf{a} + l_2\mathbf{a} \cdot \mathbf{b} + l_3\mathbf{a} \cdot \mathbf{c} = l_1$, etc.

For example, the unit vectors **i**, **j**, **k** in (8), Sec. 9.1, associated with a Cartesian coordinate system form an orthonormal basis, called the **standard basis** with respect to the given coordinate system. ■

EXAMPLE 5 Orthogonal Straight Lines in the Plane

Find the straight line L_1 through the point P: (1, 3) in the xy-plane and perpendicular to the straight line $L_2 : x - 2y + 2 = 0$; see Fig. 183.

Solution. The idea is to write a general straight line $L_1 : a_1 x + a_2 y = c$ as $\mathbf{a} \cdot \mathbf{r} = c$ with $\mathbf{a} = [a_1, a_2] \neq \mathbf{0}$ and $\mathbf{r} = [x, y]$, according to (2). Now the line L_1^* through the origin and parallel to L_1 is $\mathbf{a} \cdot \mathbf{r} = 0$. Hence, by Theorem 1, the vector $\mathbf{a}$ is perpendicular to $\mathbf{r}$. Hence it is perpendicular to L_1^* and also to L_1 because L_1 and L_1^* are parallel. $\mathbf{a}$ is called a **normal vector** of L_1 (and of L_1^*).

Now a normal vector of the given line $x - 2y + 2 = 0$ is $\mathbf{b} = [1, -2]$. Thus L_1 is perpendicular to L_2 if $\mathbf{b} \cdot \mathbf{a} = a_1 - 2a_2 = 0$, for instance, if $\mathbf{a} = [2, 1]$. Hence L_1 is given by $2x + y = c$. It passes through P: (1, 3) when $2 \cdot 1 + 3 = c = 5$. *Answer:* $y = -2x + 5$. Show that the point of intersection is $(x, y) = (1.6, 1.8)$. ■

EXAMPLE 6 Normal Vector to a Plane

Find a unit vector perpendicular to the plane $4x + 2y + 4z = -7$.

Solution. Using (2), we may write any plane in space as

$$\mathbf{a} \cdot \mathbf{r} = a_1 x + a_2 y + a_3 z = c \tag{13}$$

where $\mathbf{a} = [a_1, a_2, a_3] \neq \mathbf{0}$ and $\mathbf{r} = [x, y, z]$. The unit vector in the direction of $\mathbf{a}$ is (Fig. 184)

$$\mathbf{n} = \frac{1}{|\mathbf{a}|}\mathbf{a}.$$

Dividing by $|\mathbf{a}|$, we obtain from (13)

$$\mathbf{n} \cdot \mathbf{r} = p \qquad \text{where} \qquad p = \frac{c}{|\mathbf{a}|}. \tag{14}$$

From (12) we see that p is the projection of $\mathbf{r}$ in the direction of $\mathbf{n}$. This projection has the same constant value $c/|\mathbf{a}|$ for the position vector $\mathbf{r}$ of any point in the plane. Clearly this holds if and only if $\mathbf{n}$ is perpendicular to the plane. $\mathbf{n}$ is called a **unit normal vector** of the plane (the other being $-\mathbf{n}$).

Furthermore, from this and the definition of projection, it follows that $|p|$ is the distance of the plane from the origin. Representation (14) is called **Hesse's[2] normal form** of a plane. In our case, $\mathbf{a} = [4, 2, 4]$, $c = -7$, $|\mathbf{a}| = 6$, $\mathbf{n} = \frac{1}{6}\mathbf{a} = [\frac{2}{3}, \frac{1}{3}, \frac{2}{3}]$, and the plane has the distance $\frac{7}{6}$ from the origin. ■

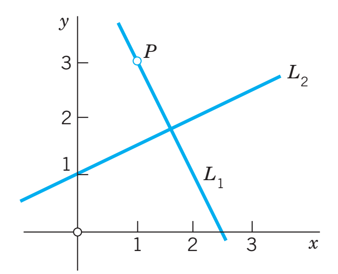


Fig. 183. Example 5

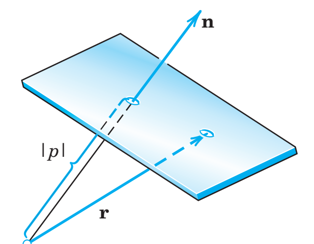


Fig. 184. Normal vector to a plane

[2] LUDWIG OTTO HESSE (1811–1874), German mathematician who contributed to the theory of curves and surfaces.

PROBLEM SET 9.2

1–10 INNER PRODUCT

Let $\mathbf{a} = [1, -3, 5]$, $\mathbf{b} = [4, 0, 8]$, $\mathbf{c} = [-2, 9, 1]$. Find:

1. $\mathbf{a} \cdot \mathbf{b}$, $\mathbf{b} \cdot \mathbf{a}$, $\mathbf{b} \cdot \mathbf{c}$

2. $(-3\mathbf{a} + 5\mathbf{c}) \cdot \mathbf{b}$, $15(\mathbf{a} - \mathbf{c}) \cdot \mathbf{b}$

3. $|\mathbf{a}|$, $|2\mathbf{b}|$, $|-\mathbf{c}|$

4. $|\mathbf{a} + \mathbf{b}|$, $|\mathbf{a}| + |\mathbf{b}|$

5. $|\mathbf{b} + \mathbf{c}|$, $|\mathbf{b}| + |\mathbf{c}|$

6. $|\mathbf{a} + \mathbf{c}|^2 + |\mathbf{a} - \mathbf{c}|^2 - 2(|\mathbf{a}|^2 + |\mathbf{c}|^2)$

7. $|\mathbf{a} \cdot \mathbf{c}|$, $|\mathbf{a}||\mathbf{c}|$

8. $5\mathbf{a} \cdot 13\mathbf{b}$, $65\mathbf{a} \cdot \mathbf{b}$

9. $15\mathbf{a} \cdot \mathbf{b} + 15\mathbf{a} \cdot \mathbf{c}$, $15\mathbf{a} \cdot (\mathbf{b} + \mathbf{c})$

10. $\mathbf{a} \cdot (\mathbf{b} - \mathbf{c})$, $(\mathbf{a} - \mathbf{b}) \cdot \mathbf{c}$

11–16 GENERAL PROBLEMS

11. What laws do Probs. 1 and 4–7 illustrate?

12. What does $\mathbf{u} \cdot \mathbf{v} = \mathbf{u} \cdot \mathbf{w}$ imply if $\mathbf{u} = \mathbf{0}$? If $\mathbf{u} \neq \mathbf{0}$?

13. Prove the Cauchy–Schwarz inequality.

14. Verify the Cauchy–Schwarz and triangle inequalities for the above **a** and **b**.

15. Prove the parallelogram equality. Explain its name.

16. Triangle inequality. Prove Eq. (7). *Hint.* Use Eq. (3) for $|\mathbf{a} + \mathbf{b}|$ and Eq. (6) to prove the square of Eq. (7), then take roots.

17–20 WORK

Find the work done by a force **p** acting on a body if the body is displaced along the straight segment $\overline{AB}$ from A to B. Sketch $\overline{AB}$ and **p**. Show the details.

17. $\mathbf{p} = [2, 5, 0]$, A: (1, 3, 3), B: (3, 5, 5)

18. $\mathbf{p} = [-1, -2, 4]$, A: (0, 0, 0), B: (6, 7, 5)

19. $\mathbf{p} = [0, 4, 3]$, A: (4, 5, −1), B: (1, 3, 0)

20. $\mathbf{p} = [6, -3, -3]$, A: (1, 5, 2), B: (3, 4, 1)

21. Resultant. Is the work done by the resultant of two forces in a displacement the sum of the work done by each of the forces separately? Give proof or counterexample.

22–30 ANGLE BETWEEN VECTORS

Let $\mathbf{a} = [1, 1, 0]$, $\mathbf{b} = [3, 2, 1]$, and $\mathbf{c} = [1, 0, 2]$. Find the angle between:

22. **a**, **b**

23. **b**, **c**

24. $\mathbf{a} + \mathbf{c}$, $\mathbf{b} + \mathbf{c}$

25. What will happen to the angle in Prob. 24 if we replace **c** by $n\mathbf{c}$ with larger and larger n?

26. Cosine law. Deduce the law of cosines by using vectors **a**, **b**, and $\mathbf{a} - \mathbf{b}$.

27. Addition law. $\cos(\alpha - \beta) = \cos\alpha \cos\beta + \sin\alpha \sin\beta$. Obtain this by using $\mathbf{a} = [\cos\alpha, \sin\alpha]$, $\mathbf{b} = [\cos\beta, \sin\beta]$ where $0 \leqq \alpha \leqq \beta \leqq 2\pi$.

28. Triangle. Find the angles of the triangle with vertices A: (0, 0, 2), B: (3, 0, 2), and C: (1, 1, 1). Sketch the triangle.

29. Parallelogram. Find the angles if the vertices are (0, 0), (6, 0), (8, 3), and (2, 3).

30. Distance. Find the distance of the point A: (1, 0, 2) from the plane P: $3x + y + z = 9$. Make a sketch.

31–35 ORTHOGONALITY is particularly important, mainly because of orthogonal coordinates, such as ***Cartesian coordinates***, whose ***natural basis*** [Eq. (9), Sec. 9.1], consists of three orthogonal unit vectors.

31. For what values of a_1 are $[a_1, 4, 3]$ and $[3, -2, 12]$ orthogonal?

32. Planes. For what c are $3x + z = 5$ and $8x - y + cz = 9$ orthogonal?

33. Unit vectors. Find all unit vectors $\mathbf{a} = [a_1, a_2]$ in the plane orthogonal to [4, 3].

34. Corner reflector. Find the angle between a light ray and its reflection in three orthogonal plane mirrors, known as *corner reflector*.

35. Parallelogram. When will the diagonals be orthogonal? Give a proof.

36–40 COMPONENT IN THE DIRECTION OF A VECTOR

Find the component of **a** in the direction of **b**. Make a sketch.

36. $\mathbf{a} = [1, 1, 1]$, $\mathbf{b} = [2, 1, 3]$

37. $\mathbf{a} = [3, 4, 0]$, $\mathbf{b} = [4, -3, 2]$

38. $\mathbf{a} = [8, 2, 0]$, $\mathbf{b} = [-4, -1, 0]$

39. When will the component (the projection) of **a** in the direction of **b** be equal to the component (the projection) of **b** in the direction of **a**? First guess.

40. What happens to the component of **a** in the direction of **b** if you change the length of **b**?

9.3 Vector Product (Cross Product)

We shall define another form of multiplication of vectors, inspired by applications, whose result will be a *vector*. This is in contrast to the dot product of Sec. 9.2 where multiplication resulted in a *scalar*. We can construct a vector $\mathbf{v}$ that is perpendicular to two vectors $\mathbf{a}$ and $\mathbf{b}$, which are two sides of a parallelogram on a plane in space as indicated in Fig. 185, such that the length $|\mathbf{v}|$ is numerically equal to the area of that parallelogram. Here then is the new concept.

DEFINITION

Vector Product (Cross Product, Outer Product) of Vectors

The **vector product** or **cross product** $\mathbf{a} \times \mathbf{b}$ (read "**a** cross **b**") of two vectors $\mathbf{a}$ and $\mathbf{b}$ is the vector $\mathbf{v}$ denoted by

$$\mathbf{v} = \mathbf{a} \times \mathbf{b}$$

I. If $\mathbf{a} = \mathbf{0}$ or $\mathbf{b} = \mathbf{0}$, then we define $\mathbf{v} = \mathbf{a} \times \mathbf{b} = \mathbf{0}$.

II. If both vectors are nonzero vectors, then vector $\mathbf{v}$ has the length

$$|\mathbf{v}| = |\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin \gamma, \tag{1}$$

where γ is the angle between $\mathbf{a}$ and $\mathbf{b}$ as in Sec. 9.2.

Furthermore, by design, $\mathbf{a}$ and $\mathbf{b}$ form the sides of a parallelogram on a plane in space. The parallelogram is shaded in blue in Fig. 185. The area of this blue parallelogram is precisely given by Eq. (1), so that the length $|\mathbf{v}|$ of the vector $\mathbf{v}$ is equal to the area of that parallelogram.

III. If $\mathbf{a}$ and $\mathbf{b}$ lie in the same straight line, i.e., $\mathbf{a}$ and $\mathbf{b}$ have the same or opposite directions, then γ is $0°$ or $180°$ so that $\sin \gamma = 0$. In that case $|\mathbf{v}| = 0$ so that $\mathbf{v} = \mathbf{a} \times \mathbf{b} = \mathbf{0}$.

IV. If cases I and III do not occur, then $\mathbf{v}$ is a nonzero vector. The direction of $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a}$, $\mathbf{b}$, $\mathbf{v}$—precisely in this order (!)—form a right-handed triple as shown in Figs. 185–187 and explained below.

Another term for vector product is outer product.

Remark. Note that I and III completely characterize the exceptional case when the cross product is equal to the zero vector, and II and IV the regular case where the cross product is perpendicular to two vectors.

Just as we did with the dot product, we would also like to express the cross product in components. Let $\mathbf{a} = [a_1, a_2, a_3]$ and $\mathbf{b} = [b_1, b_2, b_3]$. Then $\mathbf{v} = [v_1, v_2, v_3] = \mathbf{a} \times \mathbf{b}$ has the components

$$v_1 = a_2b_3 - a_3b_2, \qquad v_2 = a_3b_1 - a_1b_3, \qquad v_3 = a_1b_2 - a_2b_1. \tag{2}$$

Here the Cartesian coordinate system is *right-handed*, as explained below (see also Fig. 188). (For a left-handed system, each component of $\mathbf{v}$ must be multiplied by -1. Derivation of (2) in App. 4.)

Right-Handed Triple. A triple of vectors **a**, **b**, **v** is *right-handed* if the vectors in the given order assume the same sort of orientation as the thumb, index finger, and middle finger of the right hand when these are held as in Fig. 186. We may also say that if **a** is rotated into the direction of **b** through the angle $\gamma\,(<\pi)$, then **v** advances in the same direction as a right-handed screw would if turned in the same way (Fig. 187).

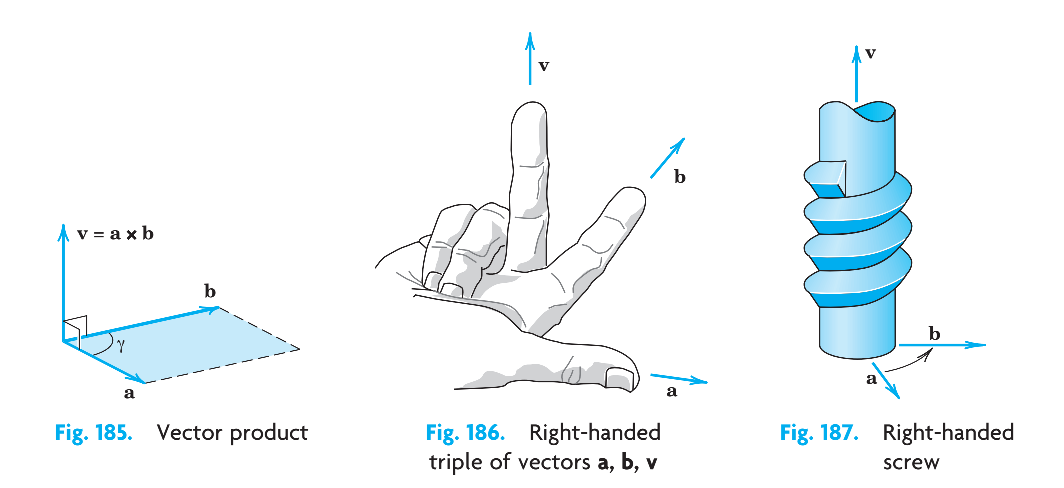


Fig. 185. Vector product

Fig. 186. Right-handed triple of vectors **a, b, v**

Fig. 187. Right-handed screw

Right-Handed Cartesian Coordinate System. The system is called **right-handed** if the corresponding unit vectors **i**, **j**, **k** in the positive directions of the axes (see Sec. 9.1) form a right-handed triple as in Fig. 188a. The system is called **left-handed** if the sense of **k** is reversed, as in Fig. 188b. In applications, we prefer right-handed systems.

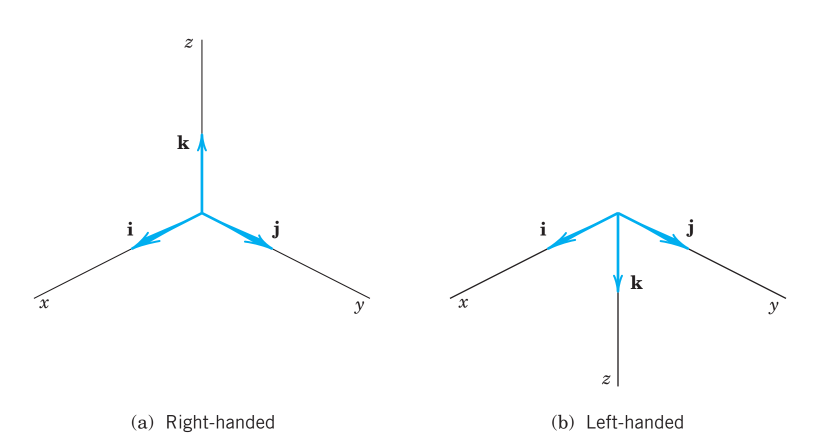


Fig. 188. The two types of Cartesian coordinate systems

How to Memorize (2). If you know second- and third-order determinants, you see that (2) can be written

$$(2^*) \qquad v_1 = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix}, \qquad v_2 = -\begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} = +\begin{vmatrix} a_3 & a_1 \\ b_3 & b_1 \end{vmatrix}, \qquad v_3 = \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix}$$

and $\mathbf{v} = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ is the expansion of the following symbolic determinant by its first row. (We call the determinant "symbolic" because the first row consists of vectors rather than of numbers.)

$$(2^{**}) \qquad \mathbf{v} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} = \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} \mathbf{i} - \begin{vmatrix} a_1 & a_3 \\ b_1 & b_3 \end{vmatrix} \mathbf{j} + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_2 \end{vmatrix} \mathbf{k}.$$

For a left-handed system the determinant has a minus sign in front.

EXAMPLE 1 **Vector Product**

For the vector product $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ of $\mathbf{a} = [1, 1, 0]$ and $\mathbf{b} = [3, 0, 0]$ in right-handed coordinates we obtain from (2)

$$v_1 = 0, \qquad v_2 = 0, \qquad v_3 = 1 \cdot 0 - 1 \cdot 3 = -3.$$

We confirm this by (2**):

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 1 & 1 & 0 \\ 3 & 0 & 0 \end{vmatrix} = \begin{vmatrix} 1 & 0 \\ 0 & 0 \end{vmatrix} \mathbf{i} - \begin{vmatrix} 1 & 0 \\ 3 & 0 \end{vmatrix} \mathbf{j} + \begin{vmatrix} 1 & 1 \\ 3 & 0 \end{vmatrix} \mathbf{k} = -3\mathbf{k} = [0, 0, -3].$$

To check the result in this simple case, sketch **a**, **b**, and **v**. Can you see that two vectors in the *xy*-plane must always have their vector product parallel to the *z*-axis (or equal to the zero vector)? ■

EXAMPLE 2 **Vector Products of the Standard Basis Vectors**

$$(3) \qquad \begin{aligned} \mathbf{i} \times \mathbf{j} &= \mathbf{k}, & \mathbf{j} \times \mathbf{k} &= \mathbf{i}, & \mathbf{k} \times \mathbf{i} &= \mathbf{j} \\ \mathbf{j} \times \mathbf{i} &= -\mathbf{k}, & \mathbf{k} \times \mathbf{j} &= -\mathbf{i}, & \mathbf{i} \times \mathbf{k} &= -\mathbf{j}. \end{aligned}$$

We shall use this in the next proof. ■

THEOREM 1 **General Properties of Vector Products**

(a) *For every scalar l,*

$$(4) \qquad (l\mathbf{a}) \times \mathbf{b} = l(\mathbf{a} \times \mathbf{b}) = \mathbf{a} \times (l\mathbf{b}).$$

(b) *Cross multiplication is distributive with respect to vector addition; that is,*

$$(5) \qquad \begin{aligned} (\alpha) \quad & \mathbf{a} \times (\mathbf{b} + \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c}), \\ (\beta) \quad & (\mathbf{a} + \mathbf{b}) \times \mathbf{c} = (\mathbf{a} \times \mathbf{c}) + (\mathbf{b} \times \mathbf{c}). \end{aligned}$$

(c) *Cross multiplication is* ***not commutative*** *but* ***anticommutative;*** *that is,*

$$(6) \qquad \mathbf{b} \times \mathbf{a} = -(\mathbf{a} \times \mathbf{b}) \qquad \text{(Fig. 189).}$$

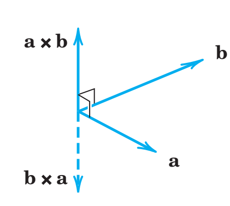


Fig. 189. Anticommutativity of cross multiplication

(d) *Cross multiplication is **not associative**; that is, in general,*

(7) $$\mathbf{a} \times (\mathbf{b} \times \mathbf{c}) \neq (\mathbf{a} \times \mathbf{b}) \times \mathbf{c}$$

so that the parentheses cannot be omitted.

PROOF Equation (4) follows directly from the definition. In (5α), formula (2*) gives for the first component on the left

$$\begin{vmatrix} a_2 & a_3 \\ b_2 + c_2 & b_3 + c_3 \end{vmatrix} = a_2(b_3 + c_3) - a_3(b_2 + c_2)$$

$$= (a_2b_3 - a_3b_2) + (a_2c_3 - a_3c_2)$$

$$= \begin{vmatrix} a_2 & a_3 \\ b_2 & b_3 \end{vmatrix} + \begin{vmatrix} a_2 & a_3 \\ c_2 & c_3 \end{vmatrix}.$$

By (2*) the sum of the two determinants is the first component of $(\mathbf{a} \times \mathbf{b}) + (\mathbf{a} \times \mathbf{c})$, the right side of (5α). For the other components in (5α) and in 5(β), equality follows by the same idea.

Anticommutativity (6) follows from (2**) by noting that the interchange of Rows 2 and 3 multiplies the determinant by -1. We can confirm this geometrically if we set $\mathbf{a} \times \mathbf{b} = \mathbf{v}$ and $\mathbf{b} \times \mathbf{a} = \mathbf{w}$; then $|\mathbf{v}| = |\mathbf{w}|$ by (1), and for $\mathbf{b}$, $\mathbf{a}$, $\mathbf{w}$ to form a *right-handed* triple, we must have $\mathbf{w} = -\mathbf{v}$.

Finally, $\mathbf{i} \times (\mathbf{i} \times \mathbf{j}) = \mathbf{i} \times \mathbf{k} = -\mathbf{j}$, whereas $(\mathbf{i} \times \mathbf{i}) \times \mathbf{j} = \mathbf{0} \times \mathbf{j} = \mathbf{0}$ (see Example 2). This proves (7). ■

Typical Applications of Vector Products

EXAMPLE 3 **Moment of a Force**

In mechanics the moment m of a force $\mathbf{p}$ about a point Q is defined as the product $m = |\mathbf{p}|d$, where d is the (perpendicular) distance between Q and the line of action L of $\mathbf{p}$ (Fig. 190). If $\mathbf{r}$ is the vector from Q to any point A on L, then $d = |\mathbf{r}| \sin \gamma$, as shown in Fig. 190, and

$$m = |\mathbf{r}||\mathbf{p}| \sin \gamma.$$

Since γ is the angle between $\mathbf{r}$ and $\mathbf{p}$, we see from (1) that $m = |\mathbf{r} \times \mathbf{p}|$. The vector

(8) $$\mathbf{m} = \mathbf{r} \times \mathbf{p}$$

is called the **moment vector** or **vector moment** of $\mathbf{p}$ about Q. Its magnitude is m. If $\mathbf{m} \neq \mathbf{0}$, its direction is that of the axis of the rotation about Q that $\mathbf{p}$ has the tendency to produce. This axis is perpendicular to both $\mathbf{r}$ and $\mathbf{p}$. ■

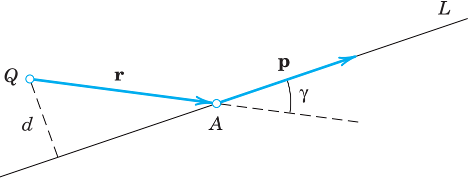


Fig. 190. Moment of a force **p**

EXAMPLE 4 Moment of a Force

Find the moment of the force **p** about the center Q of a wheel, as given in Fig. 191.

Solution. Introducing coordinates as shown in Fig. 191, we have

$$\mathbf{p} = [1000 \cos 30°, \quad 1000 \sin 30°, \quad 0] = [866, \quad 500, \quad 0], \qquad \mathbf{r} = [0, \quad 1.5, \quad 0].$$

(Note that the center of the wheel is at $y = -1.5$ on the y-axis.) Hence (8) and (2**) give

$$\mathbf{m} = \mathbf{r} \times \mathbf{p} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 1.5 & 0 \\ 866 & 500 & 0 \end{vmatrix} = 0\mathbf{i} - 0\mathbf{j} + \begin{vmatrix} 0 & 1.5 \\ 866 & 500 \end{vmatrix} \mathbf{k} = [0, 0, -1299].$$

This moment vector **m** is normal, i.e., perpendicular to the plane of the wheel. Hence it has the direction of the axis of rotation about the center Q of the wheel that the force **p** has the tendency to produce. The moment **m** points in the negative z-direction, This is, the direction in which a right-handed screw would advance if turned in that way. ■

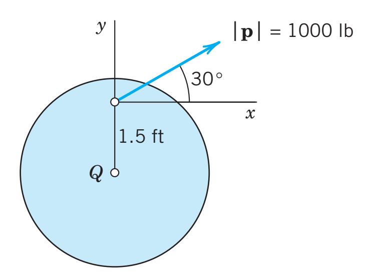


Fig. 191. Moment of a force **p**

EXAMPLE 5 Velocity of a Rotating Body

A rotation of a rigid body B in space can be simply and uniquely described by a vector **w** as follows. The direction of **w** is that of the axis of rotation and such that the rotation appears clockwise if one looks from the initial point of **w** to its terminal point. The length of **w** is equal to the **angular speed** $\omega(>0)$ of the rotation, that is, the linear (or tangential) speed of a point of B divided by its distance from the axis of rotation.

Let P be any point of B and d its distance from the axis. Then P has the speed ωd. Let **r** be the position vector of P referred to a coordinate system with origin 0 on the axis of rotation. Then $d = |\mathbf{r}| \sin \gamma$, where γ is the angle between **w** and **r**. Therefore,

$$\omega d = |\mathbf{w}||\mathbf{r}| \sin \gamma = |\mathbf{w} \times \mathbf{r}|.$$

From this and the definition of vector product we see that the velocity vector **v** of P can be represented in the form (Fig. 192)

$$\mathbf{v} = \mathbf{w} \times \mathbf{r}. \tag{9}$$

This simple formula is useful for determining **v** at any point of B. ■

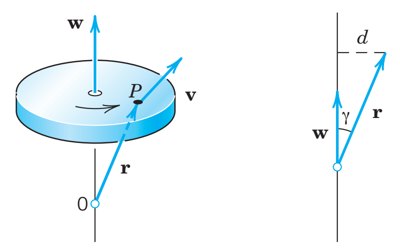


Fig. 192. Rotation of a rigid body

Scalar Triple Product

Certain products of vectors, having three or more factors, occur in applications. The most important of these products is the scalar triple product or mixed product of three vectors **a**, **b**, **c**.

(10*) $$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}).$$

The scalar triple product is indeed a scalar since (10*) involves a dot product, which in turn is a scalar. We want to express the scalar triple product in components and as a third-order determinant. To this end, let $\mathbf{a} = [a_1, a_2, a_3]$, $\mathbf{b} = [b_1, b_2, b_3]$, and $\mathbf{c} = [c_1, c_2, c_3]$. Also set $\mathbf{b} \times \mathbf{c} = \mathbf{v} = [v_1, v_2, v_3]$. Then from the dot product in components [formula (2) in Sec. 9.2] and from (2*) with **b** and **c** instead of **a** and **b** we first obtain

$$\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \mathbf{a} \cdot \mathbf{v} = a_1 v_1 + a_2 v_2 + a_3 v_3$$

$$= a_1 \begin{vmatrix} b_2 & b_3 \\ c_2 & c_3 \end{vmatrix} + a_2 \begin{vmatrix} b_3 & b_1 \\ c_3 & c_1 \end{vmatrix} + a_3 \begin{vmatrix} b_1 & b_2 \\ c_1 & c_2 \end{vmatrix}.$$

The sum on the right is the expansion of a third-order determinant by its first row. Thus we obtain the desired formula for the scalar triple product, that is,

(10) $$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = \begin{vmatrix} a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \\ c_1 & c_2 & c_3 \end{vmatrix}.$$

The most important properties of the scalar triple product are as follows.

THEOREM 2

Properties and Applications of Scalar Triple Products

(a) *In* (10) *the dot and cross can be interchanged:*

(11) $$(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}.$$

(b) Geometric interpretation. *The absolute value* $|(\mathbf{a}\ \ \mathbf{b}\ \ \mathbf{c})|$ *of* (10) *is the volume of the parallelepiped* (oblique box) *with* **a**, **b**, **c** *as edge vectors* (Fig. 193).

(c) Linear independence. *Three vectors in* R^3 *are linearly independent if and only if their scalar triple product is not zero.*

PROOF **(a)** Dot multiplication is commutative, so that by (10)

$$(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c} = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) = \begin{vmatrix} c_1 & c_2 & c_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

From this we obtain the determinant in (10) by interchanging Rows 1 and 2 and in the result Rows 2 and 3. But this does not change the value of the determinant because each interchange produces a factor -1, and $(-1)(-1) = 1$. This proves (11).

(b) The volume of that box equals the height $h = |\mathbf{a}||\cos\gamma|$ (Fig. 193) times the area of the base, which is the area $|\mathbf{b} \times \mathbf{c}|$ of the parallelogram with sides **b** and **c**. Hence the volume is

$$|\mathbf{a}||\mathbf{b} \times \mathbf{c}|\,|\cos\gamma| = |\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})| \qquad \text{(Fig. 193)}$$

as given by the absolute value of (11).

(c) Three nonzero vectors, whose initial points coincide, are linearly independent if and only if the vectors do not lie in the same plane nor lie on the same straight line.

This happens if and only if the triple product in (b) is not zero, so that the independence criterion follows. (The case of one of the vectors being the zero vector is trivial.)

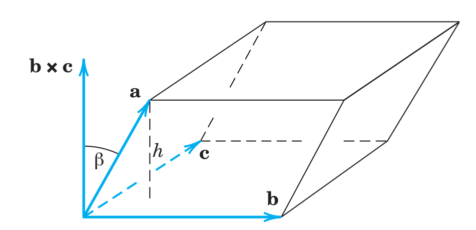


Fig. 193. Geometric interpretation of a scalar triple product

EXAMPLE 6 Tetrahedron

A tetrahedron is determined by three edge vectors **a**, **b**, **c**, as indicated in Fig. 194. Find the volume of the tetrahedron in Fig. 194, when $\mathbf{a} = [2, 0, 3]$, $\mathbf{b} = [0, 4, 1]$, $c = [5, 6, 0]$.

Solution. The volume V of the parallelepiped with these vectors as edge vectors is the absolute value of the scalar triple product

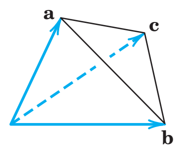


Fig. 194. Tetrahedron

$$(\mathbf{a}\quad \mathbf{b}\quad \mathbf{c}) = \begin{vmatrix} 2 & 0 & 3 \\ 0 & 4 & 1 \\ 5 & 6 & 0 \end{vmatrix} = 2\begin{vmatrix} 4 & 1 \\ 6 & 0 \end{vmatrix} + 3\begin{vmatrix} 0 & 4 \\ 5 & 6 \end{vmatrix} = -12 - 60 = -72.$$

Hence $V = 72$. The minus sign indicates that if the coordinates are right-handed, the triple **a**, **b**, **c** is left-handed. The volume of a tetrahedron is $\frac{1}{6}$ of that of the parallelepiped (can you prove it?), hence 12.

Can you sketch the tetrahedron, choosing the origin as the common initial point of the vectors? What are the coordinates of the four vertices? ■

This is the end of vector *algebra* (in space R^3 and in the plane). Vector *calculus* (differentiation) begins in the next section.

PROBLEM SET 9.3

1–10 GENERAL PROBLEMS

1. Give the details of the proofs of Eqs. (4) and (5).

2. What does $\mathbf{a} \times \mathbf{b} = \mathbf{a} \times \mathbf{c}$ with $\mathbf{a} \neq \mathbf{0}$ imply?

3. Give the details of the proofs of Eqs. (6) and (11).

4. Lagrange's identity for $|\mathbf{a} \times \mathbf{b}|$. Verify it for $\mathbf{a} = [3, 4, 2]$ and $\mathbf{b} = [1, 0, 2]$. Prove it, using $\sin^2\gamma = 1 - \cos^2\gamma$. The identity is

$$(12) \qquad |\mathbf{a} \times \mathbf{b}| = \sqrt{(\mathbf{a} \cdot \mathbf{a})\,(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2}.$$

5. What happens in Example 3 of the text if you replace $\mathbf{p}$ by $-\mathbf{p}$?

6. What happens in Example 5 if you choose a P at distance $2d$ from the axis of rotation?

7. Rotation. A wheel is rotating about the y-axis with angular speed $\omega = 20 \text{ sec}^{-1}$. The rotation appears clockwise if one looks from the origin in the positive y-direction. Find the velocity and speed at the point $[8, 6, 0]$. Make a sketch.

8. Rotation. What are the velocity and speed in Prob. 7 at the point $(4, 2, -2)$ if the wheel rotates about the line $y = x, z = 0$ with $\omega = 10 \text{ sec}^{-1}$?

9. Scalar triple product. What does $(\mathbf{a}\ \mathbf{b}\ \mathbf{c}) = 0$ imply with respect to these vectors?

10. WRITING REPORT. Summarize the most important applications discussed in this section. Give examples. No proofs.

11–23 VECTOR AND SCALAR TRIPLE PRODUCTS

With respect to right-handed Cartesian coordinates, let $\mathbf{a} = [2, 1, 0]$, $\mathbf{b} = [-3, 2, 0]$, $\mathbf{c} = [1, 4, -2]$, and $\mathbf{d} = [5, -1, 3]$. Showing details, find:

11. $\mathbf{a} \times \mathbf{b}$, $\mathbf{b} \times \mathbf{a}$, $\mathbf{a} \bullet \mathbf{b}$

12. $3\mathbf{c} \times 5\mathbf{d}$, $15\mathbf{d} \times \mathbf{c}$, $15\mathbf{d} \bullet \mathbf{c}$, $15\mathbf{c} \bullet \mathbf{d}$

13. $\mathbf{c} \times (\mathbf{a} + \mathbf{b})$, $\mathbf{a} \times \mathbf{c} + \mathbf{b} \times \mathbf{c}$

14. $4\mathbf{b} \times 3\mathbf{c} + 12\mathbf{c} \times \mathbf{b}$

15. $(\mathbf{a} + \mathbf{d}) \times (\mathbf{d} + \mathbf{a})$

16. $(\mathbf{b} \times \mathbf{c}) \bullet \mathbf{d}$, $\mathbf{b} \bullet (\mathbf{c} \times \mathbf{d})$

17. $(\mathbf{b} \times \mathbf{c}) \times \mathbf{d}$, $\mathbf{b} \times (\mathbf{c} \times \mathbf{d})$

18. $(\mathbf{a} \times \mathbf{b}) \times \mathbf{a}$, $\mathbf{a} \times (\mathbf{b} \times \mathbf{a})$

19. $(\mathbf{i}\ \mathbf{j}\ \mathbf{k})$, $(\mathbf{i}\ \mathbf{k}\ \mathbf{j})$

20. $(\mathbf{a} \times \mathbf{b}) \times (\mathbf{c} \times \mathbf{d})$, $(\mathbf{a}\ \mathbf{b}\ \mathbf{d})\mathbf{c} - (\mathbf{a}\ \mathbf{b}\ \mathbf{c})\mathbf{d}$

21. $4\mathbf{b} \times 3\mathbf{c}$, $12|\mathbf{b} \times \mathbf{c}|$, $12|\mathbf{c} \times \mathbf{b}|$

22. $(\mathbf{a} - \mathbf{b}\ \ \mathbf{c} - \mathbf{b}\ \ \mathbf{d} - \mathbf{b})$, $(\mathbf{a}\ \mathbf{c}\ \mathbf{d})$

23. $\mathbf{b} \times \mathbf{b}$, $(\mathbf{b} - \mathbf{c}) \times (\mathbf{c} - \mathbf{b})$, $\mathbf{b} \bullet \mathbf{b}$

24. TEAM PROJECT. Useful Formulas for Three and Four Vectors. Prove (13)–(16), which are often useful in practical work, and illustrate each formula with two examples. *Hint.* For (13) choose Cartesian coordinates such that $\mathbf{d} = [d_1, 0, 0]$ and $\mathbf{c} = [c_1, c_2, 0]$. Show that each side of (13) then equals $[-b_2c_2d_1, b_1c_2d_1, 0]$, and give reasons why the two sides are then equal in any Cartesian coordinate system. For (14) and (15) use (13).

(13) $\mathbf{b} \times (\mathbf{c} \times \mathbf{d}) = (\mathbf{b} \bullet \mathbf{d})\mathbf{c} - (\mathbf{b} \bullet \mathbf{c})\mathbf{d}$

(14) $(\mathbf{a} \times \mathbf{b}) \times (\mathbf{c} \times \mathbf{d}) = (\mathbf{a}\ \mathbf{b}\ \mathbf{d})\mathbf{c} - (\mathbf{a}\ \mathbf{b}\ \mathbf{c})\mathbf{d}$

(15) $(\mathbf{a} \times \mathbf{b}) \bullet (\mathbf{c} \times \mathbf{d}) = (\mathbf{a} \bullet \mathbf{c})(\mathbf{b} \bullet \mathbf{d}) - (\mathbf{a} \bullet \mathbf{d})(\mathbf{b} \bullet \mathbf{c})$

(16) $(\mathbf{a}\ \mathbf{b}\ \mathbf{c}) = (\mathbf{b}\ \mathbf{c}\ \mathbf{a}) = (\mathbf{c}\ \mathbf{a}\ \mathbf{b})$
$= -(\mathbf{c}\ \mathbf{b}\ \mathbf{a}) = -(\mathbf{a}\ \mathbf{c}\ \mathbf{b})$

25–35 APPLICATIONS

25. Moment m of a force p. Find the moment vector $\mathbf{m}$ and m of $\mathbf{p} = [2, 3, 0]$ about Q: $(2, 1, 0)$ acting on a line through A: $(0, 3, 0)$. Make a sketch.

26. Moment. Solve Prob. 25 if $\mathbf{p} = [1, 0, 3]$, Q: $(2, 0, 3)$, and A: $(4, 3, 5)$.

27. Parallelogram. Find the area if the vertices are $(4, 2, 0)$, $(10, 4, 0)$, $(5, 4, 0)$, and $(11, 6, 0)$. Make a sketch.

28. A remarkable parallelogram. Find the area of the quadrangle Q whose vertices are the midpoints of the sides of the quadrangle P with vertices A: $(2, 1, 0)$, B: $(5, -1. 0)$, C: $(8, 2, 0)$, and D: $(4, 3, 0)$. Verify that Q is a parallelogram.

29. Triangle. Find the area if the vertices are $(0, 0, 1)$, $(2, 0, 5)$, and $(2, 3, 4)$.

30. Plane. Find the plane through the points A: $(1, 2, \frac{1}{4})$, B: $(4, 2, -2)$, and C: $(0, 8, 4)$.

31. Plane. Find the plane through $(1, 3, 4)$, $(1, -2, 6)$, and $(4, 0, 7)$.

32. Parallelepiped. Find the volume if the edge vectors are $\mathbf{i} + \mathbf{j}$, $-2\mathbf{i} + 2\mathbf{k}$, and $-2\mathbf{i} - 3\mathbf{k}$. Make a sketch.

33. Tetrahedron. Find the volume if the vertices are $(1, 1, 1)$, $(5, -7, 3)$, $(7, 4, 8)$, and $(10, 7, 4)$.

34. Tetrahedron. Find the volume if the vertices are $(1, 3, 6)$, $(3, 7, 12)$, $(8, 8, 9)$, and $(2, 2, 8)$.

35. WRITING PROJECT. Applications of Cross Products. Summarize the most important applications we have discussed in this section and give a few simple examples. No proofs.

9.4 Vector and Scalar Functions and Their Fields. Vector Calculus: Derivatives

Our discussion of vector calculus begins with identifying the two types of functions on which it operates. Let P be any point in a domain of definition. Typical domains in applications are three-dimensional, or a surface or a curve in space. Then we define a **vector function** $\mathbf{v}$, whose values are vectors, that is,

$$\mathbf{v} = \mathbf{v}(P) = [v_1(P), v_2(P), v_3(P)]$$

that depends on points P in space. We say that a vector function defines a **vector field** in a domain of definition. Typical domains were just mentioned. Examples of vector fields are the field of tangent vectors of a curve (shown in Fig. 195), normal vectors of a surface (Fig. 196), and velocity field of a rotating body (Fig. 197). Note that vector functions may also depend on time t or on some other parameters.

Similarly, we define a **scalar function** f, whose values are scalars, that is,

$$f = f(P)$$

that depends on P. We say that a scalar function defines a scalar field in that three-dimensional domain or surface or curve in space. Two representative examples of scalar fields are the temperature field of a body and the pressure field of the air in Earth's atmosphere. Note that scalar functions may also depend on some parameter such as time t.

Notation. If we introduce Cartesian coordinates x, y, z, then, instead of writing $\mathbf{v}(P)$ for the vector function, we can write

$$\mathbf{v}(x, y, z) = [v_1(x, y, z), \quad v_2(x, y, z), \quad v_3(x, y, z)].$$

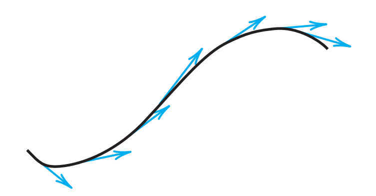


Fig. 195. Field of tangent vectors of a curve

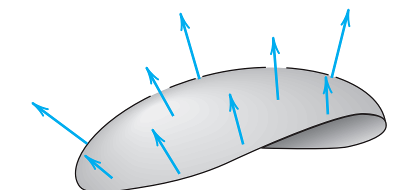


Fig. 196. Field of normal vectors of a surface

We have to keep in mind that the components depend on our choice of coordinate system, whereas a vector field that has a physical or geometric meaning should have magnitude and direction depending only on P, not on the choice of coordinate system.

Similarly, for a scalar function, we write

$$f(P) = f(x, y, z).$$

We illustrate our discussion of vector functions, scalar functions, vector fields, and scalar fields by the following three examples.

EXAMPLE 1 **Scalar Function (Euclidean Distance in Space)**

The distance $f(P)$ of any point P from a fixed point P_0 in space is a scalar function whose domain of definition is the whole space. $f(P)$ defines a scalar field in space. If we introduce a Cartesian coordinate system and P_0 has the coordinates x_0, y_0, z_0, then f is given by the well-known formula

$$f(P) = f(x, y, z) = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2}$$

where x, y, z are the coordinates of P. If we replace the given Cartesian coordinate system with another such system by translating and rotating the given system, then the values of the coordinates of P and P_0 will in general change, but $f(P)$ will have the same value as before. Hence $f(P)$ is a scalar function. The direction cosines of the straight line through P and P_0 are not scalars because their values depend on the choice of the coordinate system. ■

EXAMPLE 2 **Vector Field (Velocity Field)**

At any instant the velocity vectors $\mathbf{v}(P)$ of a rotating body B constitute a vector field, called the **velocity field** of the rotation. If we introduce a Cartesian coordinate system having the origin on the axis of rotation, then (see Example 5 in Sec. 9.3)

$$\mathbf{v}(x, y, z) = \mathbf{w} \times \mathbf{r} = \mathbf{w} \times [x, y, z] = \mathbf{w} \times (x\mathbf{i} + y\mathbf{j} + z\mathbf{k}) \tag{1}$$

where x, y, z are the coordinates of any point P of B at the instant under consideration. If the coordinates are such that the z-axis is the axis of rotation and $\mathbf{w}$ points in the positive z-direction, then $\mathbf{w} = \omega\mathbf{k}$ and

$$\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ 0 & 0 & \omega \\ x & y & z \end{vmatrix} = \omega[-y, x, 0] = \omega(-y\mathbf{i} + x\mathbf{j}).$$

An example of a rotating body and the corresponding velocity field are shown in Fig. 197. ■

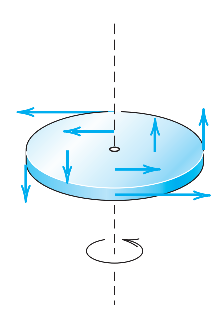


Fig. 197. Velocity field of a rotating body

EXAMPLE 3 **Vector Field (Field of Force, Gravitational Field)**

Let a particle A of mass M be fixed at a point P_0 and let a particle B of mass m be free to take up various positions P in space. Then A attracts B. According to **Newton's law of gravitation** the corresponding gravitational force $\mathbf{p}$ is directed from P to P_0, and its magnitude is proportional to $1/r^2$, where r is the distance between P and P_0, say,

$$|\mathbf{p}| = \frac{c}{r^2}, \qquad c = GMm. \tag{2}$$

Here $G = 6.67 \cdot 10^{-8}\ \text{cm}^3/(\text{g} \cdot \sec^2)$ is the gravitational constant. Hence $\mathbf{p}$ defines a vector field in space. If we introduce Cartesian coordinates such that P_0 has the coordinates x_0, y_0, z_0 and P has the coordinates x, y, z, then by the Pythagorean theorem,

$$r = \sqrt{(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2} \qquad (\geqq 0).$$

Assuming that $r > 0$ and introducing the vector

$$\mathbf{r} = [x - x_0, \quad y - y_0, \quad z - z_0] = (x - x_0)\mathbf{i} + (y - y_0)\mathbf{j} + (z - z_0)\mathbf{k},$$

we have $|\mathbf{r}| = r$, and $(-1/r)\mathbf{r}$ is a unit vector in the direction of $\mathbf{p}$; the minus sign indicates that $\mathbf{p}$ is directed from P to P_0 (Fig. 198). From this and (2) we obtain

$$\begin{aligned} \mathbf{p} = |\mathbf{p}|\left(-\frac{1}{r}\mathbf{r}\right) &= -\frac{c}{r^3}\mathbf{r} = \left[-c\,\frac{x - x_0}{r^3}, \quad -c\,\frac{y - y_0}{r^3}, \quad -c\,\frac{z - z_0}{r^3}\right] \\ &= -c\,\frac{x - x_0}{r^3}\,\mathbf{i} - c\,\frac{y - y_0}{r^3}\,\mathbf{j} - c\,\frac{z - z_0}{r^3}\,\mathbf{k}. \end{aligned} \tag{3}$$

This vector function describes the gravitational force acting on B. ■

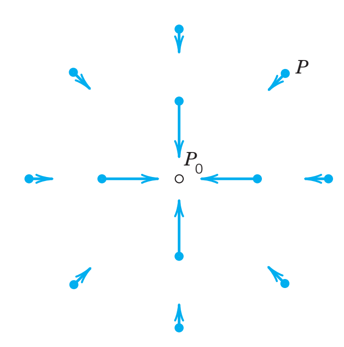


Fig. 198. Gravitational field in Example 3

Vector Calculus

The student may be pleased to learn that many of the concepts covered in (regular) calculus carry over to vector calculus. Indeed, we show how the basic concepts of convergence, continuity, and differentiability from calculus can be defined for vector functions in a simple and natural way. Most important of these is the derivative of a vector function.

Convergence. An infinite sequence of vectors $\mathbf{a}_{(n)}$, $n = 1, 2, \cdots$, is said to **converge** if there is a vector $\mathbf{a}$ such that

$$\lim_{n\to\infty} |\mathbf{a}_{(n)} - \mathbf{a}| = 0. \tag{4}$$

$\mathbf{a}$ is called the **limit vector** of that sequence, and we write

$$\lim_{n\to\infty} \mathbf{a}_{(n)} = \mathbf{a}. \tag{5}$$

If the vectors are given in Cartesian coordinates, then this sequence of vectors converges to $\mathbf{a}$ if and only if the three sequences of components of the vectors converge to the corresponding components of $\mathbf{a}$. We leave the simple proof to the student.

Similarly, a vector function $\mathbf{v}(t)$ of a real variable t is said to have the **limit** $\boldsymbol{l}$ as t approaches t_0, if $\mathbf{v}(t)$ is defined in some neighborhood of t_0 (possibly except at t_0) and

$$\lim_{t\to t_0} |\mathbf{v}(t) - \boldsymbol{l}| = 0. \tag{6}$$

Then we write

$$\lim_{t\to t_0} \mathbf{v}(t) = \boldsymbol{l}. \tag{7}$$

Here, a *neighborhood* of t_0 is an interval (segment) on the t-axis containing t_0 as an interior point (not as an endpoint).

Continuity. A vector function $\mathbf{v}(t)$ is said to be **continuous** at $t = t_0$ if it is defined in some neighborhood of t_0 (including at t_0 itself!) and

$$\lim_{t\to t_0} \mathbf{v}(t) = \mathbf{v}(t_0). \tag{8}$$

If we introduce a Cartesian coordinate system, we may write

$$\mathbf{v}(t) = [v_1(t), v_2(t), v_3(t)] = v_1(t)\mathbf{i} + v_2(t)\mathbf{j} + v_3(t)\mathbf{k}.$$

Then $\mathbf{v}(t)$ is continuous at t_0 if and only if its three components are continuous at t_0.

We now state the most important of these definitions.

DEFINITION

Derivative of a Vector Function

A vector function $\mathbf{v}(t)$ is said to be **differentiable** at a point t if the following limit exists:

$$\mathbf{v}'(t) = \lim_{\Delta t \to 0} \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t}. \tag{9}$$

This vector $\mathbf{v}'(t)$ is called the **derivative** of $\mathbf{v}(t)$. See Fig. 199.

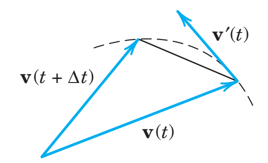


Fig. 199. Derivative of a vector function

In components with respect to a given Cartesian coordinate system,

$$\mathbf{v}'(t) = [v_1'(t), \quad v_2'(t), \quad v_3'(t)]. \tag{10}$$

Hence the derivative $\mathbf{v}'(t)$ ***is obtained by differentiating each component separately.*** For instance, if $\mathbf{v} = [t, t^2, 0]$, then $\mathbf{v}' = [1, 2t, 0]$.

Equation (10) follows from (9) and conversely because (9) is a "vector form" of the usual formula of calculus by which the derivative of a function of a single variable is defined. [The curve in Fig. 199 is the locus of the terminal points representing $\mathbf{v}(t)$ for values of the independent variable in some interval containing t and $t + \Delta t$ in (9)]. It follows that the familiar differentiation rules continue to hold for differentiating vector functions, for instance,

$$(c\mathbf{v})' = c\mathbf{v}' \qquad (c \text{ constant}),$$

$$(\mathbf{u} + \mathbf{v})' = \mathbf{u}' + \mathbf{v}'$$

and in particular

$$(\mathbf{u} \cdot \mathbf{v})' = \mathbf{u}' \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{v}' \tag{11}$$

$$(\mathbf{u} \times \mathbf{v})' = \mathbf{u}' \times \mathbf{v} + \mathbf{u} \times \mathbf{v}' \tag{12}$$

$$(\mathbf{u} \quad \mathbf{v} \quad \mathbf{w})' = (\mathbf{u}' \quad \mathbf{v} \quad \mathbf{w}) + (\mathbf{u} \quad \mathbf{v}' \quad \mathbf{w}) + (\mathbf{u} \quad \mathbf{v} \quad \mathbf{w}'). \tag{13}$$

The simple proofs are left to the student. In (12), note the order of the vectors carefully because cross multiplication is not commutative.

EXAMPLE 4 **Derivative of a Vector Function of Constant Length**

Let $\mathbf{v}(t)$ be a vector function whose length is constant, say, $|\mathbf{v}(t)| = c$. Then $|\mathbf{v}|^2 = \mathbf{v} \cdot \mathbf{v} = c^2$, and $(\mathbf{v} \cdot \mathbf{v})' = 2\mathbf{v} \cdot \mathbf{v}' = 0$, by differentiation [see (11)]. This yields the following result. *The derivative of a vector function* $\mathbf{v}(t)$ *of constant length is either the zero vector or is perpendicular to* $\mathbf{v}(t)$. ■

Partial Derivatives of a Vector Function

Our present discussion shows that partial differentiation of vector functions of two or more variables can be introduced as follows. Suppose that the components of a vector function

$$\mathbf{v} = [v_1, \quad v_2, \quad v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$$

are differentiable functions of n variables $t_1, \cdots, t_n$. Then the **partial derivative** of $\mathbf{v}$ with respect to t_m is denoted by $\partial\mathbf{v}/\partial t_m$ and is defined as the vector function

$$\frac{\partial \mathbf{v}}{\partial t_m} = \frac{\partial v_1}{\partial t_m}\mathbf{i} + \frac{\partial v_2}{\partial t_m}\mathbf{j} + \frac{\partial v_3}{\partial t_m}\mathbf{k}.$$

Similarly, second partial derivatives are

$$\frac{\partial^2 \mathbf{v}}{\partial t_l \partial t_m} = \frac{\partial^2 v_1}{\partial t_l \partial t_m}\mathbf{i} + \frac{\partial^2 v_2}{\partial t_l \partial t_m}\mathbf{j} + \frac{\partial^2 v_3}{\partial t_l \partial t_m}\mathbf{k},$$

and so on.

EXAMPLE 5 **Partial Derivatives**

Let $\mathbf{r}(t_1, t_2) = a \cos t_1 \,\mathbf{i} + a \sin t_1 \,\mathbf{j} + t_2 \,\mathbf{k}$. Then $\dfrac{\partial \mathbf{r}}{\partial t_1} = -a \sin t_1 \,\mathbf{i} + a \cos t_1 \,\mathbf{j}$ and $\dfrac{\partial \mathbf{r}}{\partial t_2} = \mathbf{k}$. ■

Various physical and geometric applications of derivatives of vector functions will be discussed in the next sections as well as in Chap. 10.

PROBLEM SET 9.4

1–8 SCALAR FIELDS IN THE PLANE

Let the temperature T in a body be independent of z so that it is given by a scalar function $T = T(x, t)$. Identify the isotherms $T(x, y) =$ const. Sketch some of them.

1. $T = x^2 - y^2$
2. $T = xy$
3. $T = 3x - 4y$
4. $T = \arctan(y/x)$
5. $T = y/(x^2 + y^2)$
6. $T = x/(x^2 + y^2)$
7. $T = 9x^2 + 4y^2$
8. CAS PROJECT. Scalar Fields in the Plane. Sketch or graph isotherms of the following fields and describe what they look like.

(a) $x^2 - 4x - y^2$
(b) $x^2y - y^3/3$
(c) $\cos x \sinh y$
(d) $\sin x \sinh y$
(e) $e^x \sin y$
(f) $e^{2x} \cos 2y$
(g) $x^4 - 6x^2y^2 + y^4$
(h) $x^2 - 2x - y^2$

9–14 SCALAR FIELDS IN SPACE

What kind of surfaces are the **level surfaces** $f(x, y, z) =$ const?

9. $f = 4x - 3y + 2z$
10. $f = 9(x^2 + y^2) + z^2$
11. $f = 5x^2 + 2y^2$
12. $f = z - \sqrt{x^2 + y^2}$
13. $f = z - (x^2 + y^2)$
14. $f = x - y^2$

15–20 VECTOR FIELDS

Sketch figures similar to Fig. 198. Try to interpet the field of **v** as a velocity field.

15. $\mathbf{v} = \mathbf{i} + \mathbf{j}$

16. $\mathbf{v} = -y\mathbf{i} + x\mathbf{j}$

17. $\mathbf{v} = x\mathbf{j}$

18. $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$

19. $\mathbf{v} = x\mathbf{i} - y\mathbf{j}$

20. $\mathbf{v} = y\mathbf{i} - x\mathbf{j}$

21. CAS PROJECT. Vector Fields. Plot by arrows:

(a) $\mathbf{v} = [x, x^2]$ **(b)** $\mathbf{v} = [1/y, 1/x]$

(c) $\mathbf{v} = [\cos x, \sin x]$ **(d)** $\mathbf{v} = e^{-(x^2+y^2)}[x, -y]$

22–25 DIFFERENTIATION

22. Find the first and second derivatives of $\mathbf{r} = [3 \cos 2t, 3 \sin 2t, 4t]$.

23. Prove (11)–(13). Give two typical examples for each formula.

24. Find the first partial derivatives of $\mathbf{v}_1 = [e^x \cos y, e^x \sin y]$ and $\mathbf{v}_2 = [\cos x \cosh y, -\sin x \sinh y]$.

25. WRITING PROJECT. Differentiation of Vector Functions. Summarize the essential ideas and facts and give examples of your own.

9.5 Curves. Arc Length. Curvature. Torsion

Vector calculus has important applications to curves (Sec. 9.5) and surfaces (to be covered in Sec. 10.5) in physics and geometry. The application of vector calculus to geometry is a field known as **differential geometry**. Differential geometric methods are applied to problems in mechanics, computer-aided as well as traditional engineering design, geodesy, geography, space travel, and relativity theory. For details, see [GenRef8] and [GenRef9] in App. 1.

Bodies that move in space form paths that may be represented by curves C. This and other applications show the need for **parametric representations** of C with **parameter** t, which may denote time or something else (see Fig. 200). A typical parametric representation is given by

$$\textbf{(1)} \qquad \mathbf{r}(t) = [x(t), \quad y(t), \quad z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}.$$

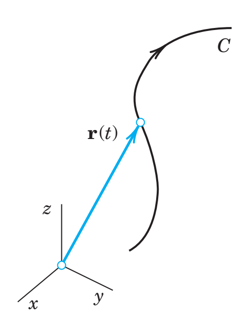


Fig. 200. Parametric representation of a curve

Here t is the parameter and x, y, z are Cartesian coordinates, that is, the usual rectangular coordinates as shown in Sec. 9.1. To each value $t = t_0$, there corresponds a point of C with position vector $\mathbf{r}(t_0)$ whose coordinates are $x(t_0), y(t_0), z(t_0)$. This is illustrated in Figs. 201 and 202.

The use of parametric representations has key advantages over other representations that involve projections into the xy-plane and xz-plane or involve a pair of equations with y or with z as independent variable. The projections look like this:

$$(2) \qquad y = f(x), \qquad z = g(x).$$

The advantages of using (1) instead of (2) are that, in (1), the coordinates x, y, z all play an equal role, that is, all three coordinates are dependent variables. Moreover, the parametric representation (1) induces an orientation on C. This means that as we increase t, we travel along the curve C in a certain direction. The sense of increasing t is called the positive sense on C. The sense of decreasing t is then called the negative sense on C, given by (1).

Examples 1–4 give parametric representations of several important curves.

EXAMPLE 1 Circle. Parametric Representation. Positive Sense

The circle $x^2 + y^2 = 4$, $z = 0$ in the xy-plane with center 0 and radius 2 can be represented parametrically by

$$\mathbf{r}(t) = [2\cos t, 2\sin t, 0] \qquad \text{or simply by} \qquad \mathbf{r}(t) = [2\cos t, 2\sin t] \qquad \text{(Fig. 201)}$$

where $0 \leqq t \leqq 2\pi$. Indeed, $x^2 + y^2 = (2\cos t)^2 + (2\sin t)^2 = 4(\cos^2 t + \sin^2 t) = 4$, For $t = 0$ we have $\mathbf{r}(0) = [2, 0]$, for $t = \frac{1}{2}\pi$ we get $\mathbf{r}(\frac{1}{2}\pi) = [0, 2]$, and so on. The positive sense induced by this representation is the counterclockwise sense.

If we replace t with $t^* = -t$, we have $t = -t^*$ and get

$$\mathbf{r}^*(t^*) = [2\cos(-t^*), 2\sin(-t^*)] = [2\cos t^*, -2\sin t^*].$$

This has reversed the orientation, and the circle is now oriented clockwise. ■

EXAMPLE 2 Ellipse

The vector function

$$\mathbf{r}(t) = [a\cos t, \quad b\sin t, \quad 0] = a\cos t\,\mathbf{i} + b\sin t\,\mathbf{j} \qquad (3) \qquad \text{(Fig. 202)}$$

represents an ellipse in the xy-plane with center at the origin and principal axes in the direction of the x- and y-axes. In fact, since $\cos^2 t + \sin^2 t = 1$, we obtain from (3)

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1, \qquad z = 0.$$

If $b = a$, then (3) represents a *circle* of radius a. ■

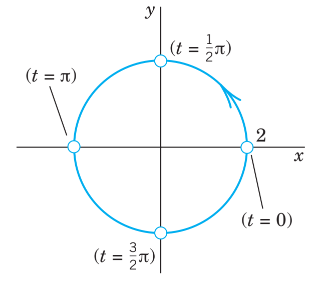


Fig. 201. Circle in Example 1

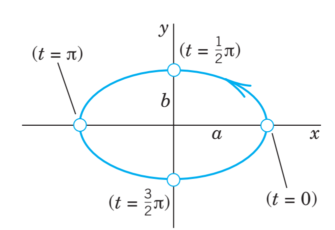


Fig. 202. Ellipse in Example 2

EXAMPLE 3 Straight Line

A straight line L through a point A with position vector $\mathbf{a}$ in the direction of a constant vector $\mathbf{b}$ (see Fig. 203) can be represented parametrically in the form

$$\mathbf{r}(t) = \mathbf{a} + t\mathbf{b} = [a_1 + tb_1, \quad a_2 + tb_2, \quad a_3 + tb_3]. \qquad (4)$$

If **b** is a unit vector, its components are the **direction cosines** of L. In this case, $|t|$ measures the distance of the points of L from A. For instance, the straight line in the xy-plane through A: (3, 2) having slope 1 is (sketch it)

$$\mathbf{r}(t) = [3,\quad 2,\quad 0] + t[1,\quad 1,\quad 0] = [3 + t,\quad 2 + t,\quad 0].$$

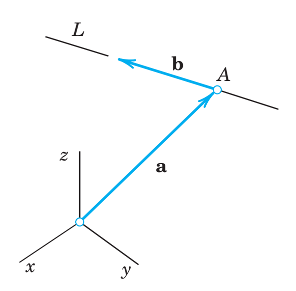


Fig. 203. Parametric representation of a straight line

A **plane curve** is a curve that lies in a plane in space. A curve that is not plane is called a **twisted curve**. A standard example of a twisted curve is the following.

EXAMPLE 4 Circular Helix

The twisted curve C represented by the vector function

$$\mathbf{r}(t) = [a \cos t,\quad a \sin t,\quad ct] = a \cos t\,\mathbf{i} + a \sin t\,\mathbf{j} + ct\,\mathbf{k} \qquad (c \neq 0) \tag{5}$$

is called a *circular helix*. It lies on the cylinder $x^2 + y^2 = a^2$. If $c > 0$, the helix is shaped like a right-handed screw (Fig. 204). If $c < 0$, it looks like a left-handed screw (Fig. 205). If $c = 0$, then (5) is a circle.

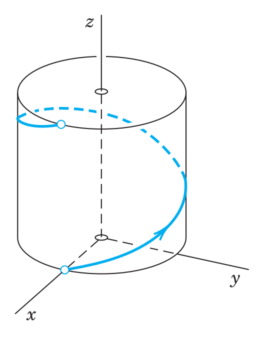


Fig. 204. Right-handed circular helix

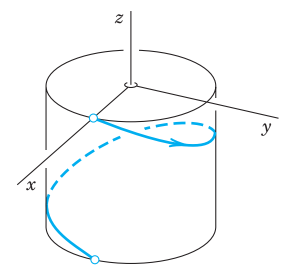


Fig. 205. Left-handed circular helix

A **simple curve** is a curve without **multiple points**, that is, without points at which the curve intersects or touches itself. Circle and helix are simple curves. Figure 206 shows curves that are not simple. An example is $[\sin 2t,\quad \cos t,\quad 0]$. Can you sketch it?

An **arc** of a curve is the portion between any two points of the curve. For simplicity, we say "curve" for curves as well as for arcs.

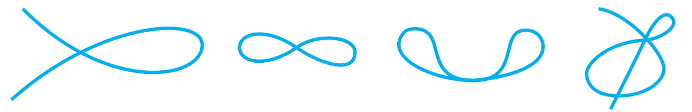


Fig. 206. Curves with multiple points

Tangent to a Curve

The next idea is the approximation of a curve by straight lines, leading to tangents and to a definition of length. Tangents are straight lines touching a curve. The **tangent** to a simple curve C at a point P of C is the limiting position of a straight line L through P and a point Q of C as Q approaches P along C. See Fig. 207.

Let us formalize this concept. If C is given by $\mathbf{r}(t)$, and P and Q correspond to t and $t + \Delta t$, then a vector in the direction of L is

$$\frac{1}{\Delta t}[\mathbf{r}(t + \Delta t) - \mathbf{r}(t)]. \tag{6}$$

In the limit this vector becomes the derivative

$$\mathbf{r}'(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}[\mathbf{r}(t + \Delta t) - \mathbf{r}(t)], \tag{7}$$

provided $\mathbf{r}(t)$ is differentiable, as we shall assume from now on. If $\mathbf{r}'(t) \neq \mathbf{0}$, we call $\mathbf{r}'(t)$ a **tangent vector** of C at P because it has the direction of the tangent. The corresponding unit vector is the **unit tangent vector** (see Fig. 207)

$$\mathbf{u} = \frac{1}{|\mathbf{r}'|}\mathbf{r}'. \tag{8}$$

Note that both $\mathbf{r}'$ and $\mathbf{u}$ point in the direction of increasing t. Hence their sense depends on the orientation of C. It is reversed if we reverse the orientation.

It is now easy to see that the **tangent** to C at P is given by

$$\mathbf{q}(w) = \mathbf{r} + w\mathbf{r}' \qquad \text{(Fig. 208).} \tag{9}$$

This is the sum of the position vector $\mathbf{r}$ of P and a multiple of the tangent vector $\mathbf{r}'$ of C at P. Both vectors depend on P. The variable w is the parameter in (9).

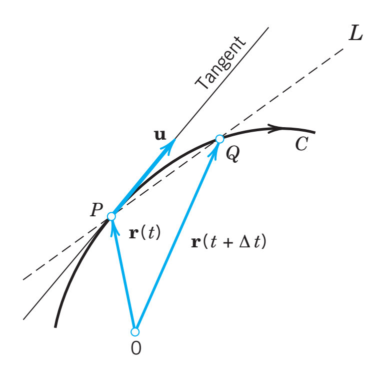


Fig. 207. Tangent to a curve

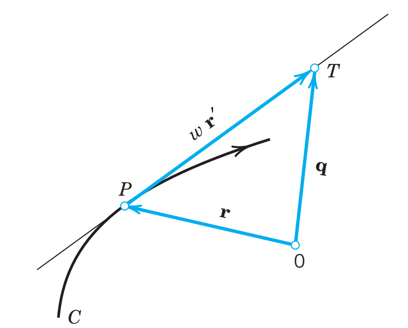


Fig. 208. Formula (9) for the tangent to a curve

EXAMPLE 5 Tangent to an Ellipse

Find the tangent to the ellipse $\frac{1}{4}x^2 + y^2 = 1$ at P: $(\sqrt{2}, 1/\sqrt{2})$.

Solution. Equation (3) with semi-axes $a = 2$ and $b = 1$ gives $\mathbf{r}(t) = [2\cos t, \sin t]$. The derivative is $\mathbf{r}'(t) = [-2\sin t, \cos t]$. Now P corresponds to $t = \pi/4$ because

$$\mathbf{r}(\pi/4) = [2\cos(\pi/4), \sin(\pi/4)] = [\sqrt{2}, \quad 1/\sqrt{2}].$$

Hence $\mathbf{r}'(\pi/4) = [-\sqrt{2},\ \ 1/\sqrt{2}]$. From (9) we thus get the *answer*

$$\mathbf{q}(w) = [\sqrt{2},\ \ 1/\sqrt{2}] + w[-\sqrt{2},\ \ 1/\sqrt{2}] = [\sqrt{2}(1-w),\ \ (1/\sqrt{2})(1+w)].$$

To check the result, sketch or graph the ellipse and the tangent. ■

Length of a Curve

We are now ready to define the length l of a curve. l will be the limit of the lengths of broken lines of n chords (see Fig. 209, where $n = 5$) with larger and larger n. For this, let $\mathbf{r}(t)$, $a \leqq t \leqq b$, represent C. For each $n = 1, 2, \cdots$, we subdivide ("partition") the interval $a \leqq t \leqq b$ by points

$$t_0 (= a), \quad t_1, \cdots, t_{n-1}, \quad t_n (= b), \qquad \text{where} \qquad t_0 < t_1 < \cdots < t_n.$$

This gives a broken line of chords with endpoints $\mathbf{r}(t_0), \cdots, \mathbf{r}(t_n)$. We do this arbitrarily but so that the greatest $|\Delta t_m| = |t_m - t_{m-1}|$ approaches 0 as $n \rightarrow \infty$. The lengths $l_1, l_2, \cdots$ of these chords can be obtained from the Pythagorean theorem. If $\mathbf{r}(t)$ has a continuous derivative $\mathbf{r}'(t)$, it can be shown that the sequence $l_1, l_2, \cdots$ has a limit, which is independent of the particular choice of the representation of C and of the choice of subdivisions. This limit is given by the integral

$$\textbf{(10)} \qquad l = \int_a^b \sqrt{\mathbf{r}' \cdot \mathbf{r}'}\, dt \qquad \left(\mathbf{r}' = \frac{d\mathbf{r}}{dt}\right).$$

l is called the **length** of C, and C is called **rectifiable**. Formula (10) is made plausible in calculus for plane curves and is proved for curves in space in [GenRef8] listed in App. 1. The actual evaluation of the integral (10) will, in general, be difficult. However, some simple cases are given in the problem set.

Arc Length s of a Curve

The length (10) of a curve C is a constant, a positive number. But if we replace the fixed b in (10) with a variable t, the integral becomes a function of t, denoted by $s(t)$ and called the *arc length function* or simply the **arc length** of C. Thus

$$\textbf{(11)} \qquad s(t) = \int_a^t \sqrt{\mathbf{r}' \cdot \mathbf{r}'}\, d\tilde{t} \qquad \left(\mathbf{r}' = \frac{d\mathbf{r}}{d\tilde{t}}\right).$$

Here the variable of integration is denoted by $\tilde{t}$ because t is now used in the upper limit.

Geometrically, $s(t_0)$ with some $t_0 > a$ is the length of the arc of C between the points with parametric values a and t_0. The choice of a (the point $s = 0$) is arbitrary; changing a means changing s by a constant.

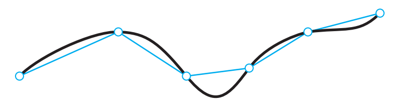


Fig. 209. Length of a curve

Linear Element *ds*. If we differentiate (11) and square, we have

$$\left(\frac{ds}{dt}\right)^2 = \frac{d\mathbf{r}}{dt} \cdot \frac{d\mathbf{r}}{dt} = |\mathbf{r}'(t)|^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 + \left(\frac{dz}{dt}\right)^2. \tag{12}$$

It is customary to write

$$d\mathbf{r} = [dx, dy, dz] = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k} \tag{13*}$$

and

$$ds^2 = d\mathbf{r} \cdot d\mathbf{r} = dx^2 + dy^2 + dz^2. \tag{13}$$

ds is called the **linear element** of C.

Arc Length as Parameter. The use of s in (1) instead of an arbitrary t simplifies various formulas. For the unit tangent vector (8) we simply obtain

$$\mathbf{u}(s) = \mathbf{r}'(s). \tag{14}$$

Indeed, $|\mathbf{r}'(s)| = (ds/ds) = 1$ in (12) shows that $\mathbf{r}'(s)$ is a unit vector. Even greater simplifications due to the use of s will occur in curvature and torsion (below).

EXAMPLE 6 Circular Helix. Circle. Arc Length as Parameter

The **helix** $\mathbf{r}(t) = [a\cos t, a\sin t, ct]$ in (5) has the derivative $\mathbf{r}'(t) = [-a\sin t, a\cos t, c]$. Hence $\mathbf{r}' \cdot \mathbf{r}' = a^2 + c^2$, a constant, which we denote by K^2. Hence the integrand in (11) is constant, equal to K, and the integral is $s = Kt$. Thus $t = s/K$, so that a representation of the helix with the arc length s as parameter is

$$\mathbf{r}^*(s) = \mathbf{r}\left(\frac{s}{K}\right) = \left[a\cos\frac{s}{K},\quad a\sin\frac{s}{K},\quad \frac{cs}{K}\right], \qquad K = \sqrt{a^2 + c^2}. \tag{15}$$

A **circle** is obtained if we set $c = 0$. Then $K = a$, $t = s/a$, and a representation with arc length s as parameter is

$$\mathbf{r}^*(s) = \mathbf{r}\left(\frac{s}{a}\right) = \left[a\cos\frac{s}{a},\quad a\sin\frac{s}{a}\right].$$

■

Curves in Mechanics. Velocity. Acceleration

Curves play a basic role in mechanics, where they may serve as paths of moving bodies. Then such a curve C should be represented by a parametric representation $\mathbf{r}(t)$ with **time** t as parameter. The tangent vector (7) of C is then called the **velocity vector v** because, being tangent, it points in the instantaneous direction of motion and its length gives the **speed** $|\mathbf{v}| = |\mathbf{r}'| = \sqrt{\mathbf{r}' \cdot \mathbf{r}'} = ds/dt$; see (12). The second derivative of $\mathbf{r}(t)$ is called the **acceleration vector** and is denoted by **a**. Its length $|\mathbf{a}|$ is called the **acceleration** of the motion. Thus

$$\mathbf{v}(t) = \mathbf{r}'(t), \qquad \mathbf{a}(t) = \mathbf{v}'(t) = \mathbf{r}''(t). \tag{16}$$

Tangential and Normal Acceleration. Whereas the velocity vector is always tangent to the path of motion, the acceleration vector will generally have another direction. We can split the acceleration vector into two directional components, that is,

$$\mathbf{a} = \mathbf{a}_{\text{tan}} + \mathbf{a}_{\text{norm}}, \tag{17}$$

where the **tangential acceleration vector** $\mathbf{a}_{\text{tan}}$ is tangent to the path (or, sometimes, $\mathbf{0}$) and the **normal acceleration vector** $\mathbf{a}_{\text{norm}}$ is normal (perpendicular) to the path (or, sometimes, $\mathbf{0}$).

Expressions for the vectors in (17) are obtained from (16) by the chain rule. We first have

$$\mathbf{v}(t) = \frac{d\mathbf{r}}{dt} = \frac{d\mathbf{r}}{ds}\frac{ds}{dt} = \mathbf{u}(s)\frac{ds}{dt}$$

where $\mathbf{u}(s)$ is the unit tangent vector (14). Another differentiation gives

$$\mathbf{a}(t) = \frac{d\mathbf{v}}{dt} = \frac{d}{dt}\left(\mathbf{u}(s)\frac{ds}{dt}\right) = \frac{d\mathbf{u}}{ds}\left(\frac{ds}{dt}\right)^2 + \mathbf{u}(s)\frac{d^2s}{dt^2}. \tag{18}$$

Since the tangent vector $\mathbf{u}(s)$ has constant length (length one), its derivative $d\mathbf{u}/ds$ is perpendicular to $\mathbf{u}(s)$, from the result in Example 4 in Sec. 9.4. Hence the first term on the right of (18) is the normal acceleration vector, and the second term on the right is the tangential acceleration vector, so that (18) is of the form (17).

Now the length $|\mathbf{a}_{\text{tan}}|$ is the absolute value of the projection of $\mathbf{a}$ in the direction of $\mathbf{v}$, given by (11) in Sec. 9.2 with $\mathbf{b} = \mathbf{v}$; that is, $|\mathbf{a}_{\text{tan}}| = |\mathbf{a} \cdot \mathbf{v}|/|\mathbf{v}|$. Hence $\mathbf{a}_{\text{tan}}$ is this expression times the unit vector $(1/|\mathbf{v}|)\mathbf{v}$ in the direction of $\mathbf{v}$, that is,

$$\mathbf{a}_{\text{tan}} = \frac{\mathbf{a} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}\mathbf{v}. \qquad \text{Also,} \qquad \mathbf{a}_{\text{norm}} = \mathbf{a} - \mathbf{a}_{\text{tan}}. \tag{18*}$$

We now turn to two examples that are relevant to applications in space travel. They deal with the *centripetal* and *centrifugal* accelerations, as well as the *Coriolis acceleration*.

EXAMPLE 7 Centripetal Acceleration. Centrifugal Force

The vector function

$$\mathbf{r}(t) = [R \cos \omega t, \quad R \sin \omega t] = R \cos \omega t \,\mathbf{i} + R \sin \omega t \,\mathbf{j} \qquad \text{(Fig. 210)}$$

(with fixed $\mathbf{i}$ and $\mathbf{j}$) represents a circle C of radius R with center at the origin of the xy-plane and describes the motion of a small body B counterclockwise around the circle. Differentiation gives the velocity vector

$$\mathbf{v} = \mathbf{r}' = [-R\omega \sin \omega t, \quad R\omega \cos \omega t] = -R\omega \sin \omega t \,\mathbf{i} + R\omega \cos \omega t \,\mathbf{j} \qquad \text{(Fig. 210)}$$

$\mathbf{v}$ is tangent to C. Its magnitude, the speed, is

$$|\mathbf{v}| = |\mathbf{r}'| = \sqrt{\mathbf{r}' \cdot \mathbf{r}'} = R\omega.$$

Hence it is constant. The speed divided by the distance R from the center is called the **angular speed**. It equals ω, so that it is constant, too. Differentiating the velocity vector, we obtain the acceleration vector

$$\mathbf{a} = \mathbf{v}' = [-R\omega^2 \cos \omega t, \quad -R\omega^2 \sin \omega t] = -R\omega^2 \cos \omega t \,\mathbf{i} - R\omega^2 \sin \omega t \,\mathbf{j}. \tag{19}$$

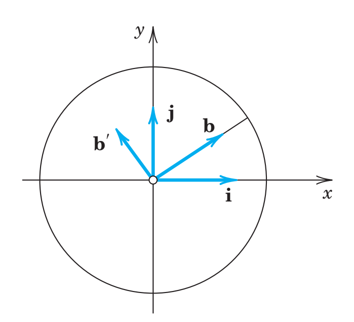


Fig. 210. Centripetal acceleration **a**

This shows that $\mathbf{a} = -\omega^2\mathbf{r}$ (Fig. 210), so that there is an acceleration toward the center, called the **centripetal acceleration** of the motion. It occurs because the velocity vector is changing direction at a constant rate. Its magnitude is constant, $|\mathbf{a}| = \omega^2|\mathbf{r}| = \omega^2 R$. Multiplying **a** by the mass m of B, we get the **centripetal force** $m\mathbf{a}$. The opposite vector $-m\mathbf{a}$ is called the **centrifugal force**. At each instant these two forces are in equilibrium.

We see that in this motion the acceleration vector is normal (perpendicular) to C; hence there is no tangential acceleration. ■

EXAMPLE 8 Superposition of Rotations. Coriolis Acceleration

A projectile is moving with constant speed along a meridian of the rotating earth in Fig. 211. Find its acceleration.

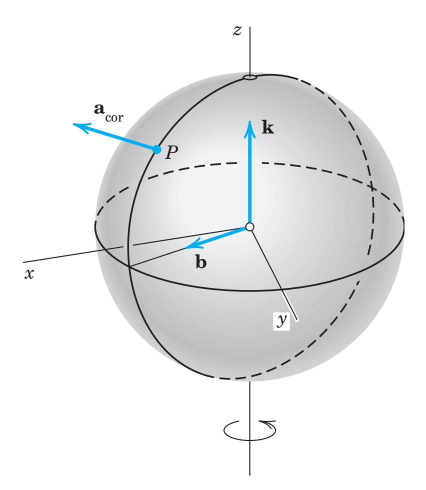


Fig. 211. Example 8. Superposition of two rotations

Solution. Let x, y, z be a fixed Cartesian coordinate system in space, with unit vectors **i**, **j**, **k** in the directions of the axes. Let the Earth, together with a unit vector **b**, be rotating about the z-axis with angular speed $\omega > 0$ (see Example 7). Since **b** is rotating together with the Earth, it is of the form

$$\mathbf{b}(t) = \cos \omega t \, \mathbf{i} + \sin \omega t \, \mathbf{j}.$$

Let the projectile be moving on the meridian whose plane is spanned by **b** and **k** (Fig. 211) with constant angular speed $\omega > 0$. Then its position vector in terms of **b** and **k** is

$$\mathbf{r}(t) = R \cos \gamma t \, \mathbf{b}(t) + R \sin \gamma t \, \mathbf{k} \qquad (R = \text{Radius of the Earth}).$$

We have finished setting up the model. Next, we apply vector calculus to obtain the desired acceleration of the projectile. Our result will be unexpected—and highly relevant for air and space travel. The first and second derivatives of $\mathbf{b}$ with respect to t are

$$\begin{aligned} \mathbf{b}'(t) &= -\omega \sin \omega t\, \mathbf{i} + \omega \cos \omega t\, \mathbf{j} \\ \mathbf{b}''(t) &= -\omega^2 \cos \omega t\, \mathbf{i} - \omega^2 \sin \omega t\, \mathbf{j} = -\omega^2 \mathbf{b}(t). \end{aligned} \tag{20}$$

The first and second derivatives of $\mathbf{r}(t)$ with respect to t are

$$\begin{aligned} \mathbf{v} = \mathbf{r}'(t) &= R \cos \gamma t\, \mathbf{b}' - \gamma R \sin \gamma t\, \mathbf{b} + \gamma R \cos \gamma t\, \mathbf{k} \\ \mathbf{a} = \mathbf{v}' &= R \cos \gamma t\, \mathbf{b}'' - 2\gamma R \sin \gamma t\, \mathbf{b}' - \gamma^2 R \cos \gamma t\, \mathbf{b} - \gamma^2 R \sin \gamma t\, \mathbf{k} \\ &= R \cos \gamma t\, \mathbf{b}'' - 2\gamma R \sin \gamma t\, \mathbf{b}' - \gamma^2 \mathbf{r}. \end{aligned} \tag{21}$$

By analogy with Example 7 and because of $\mathbf{b}'' = -\omega^2 \mathbf{b}$ in (20) we conclude that the first term in $\mathbf{a}$ (involving ω in $\mathbf{b}''$!) is the centripetal acceleration due to the rotation of the Earth. Similarly, the third term in the last line (involving γ!) is the centripetal acceleration due to the motion of the projectile on the meridian M of the rotating Earth.

The second, unexpected term $-2\gamma R \sin \gamma t\, \mathbf{b}'$ in $\mathbf{a}$ is called the **Coriolis acceleration**[3] (Fig. 211) and is due to the interaction of the two rotations. On the Northern Hemisphere, $\sin \gamma t > 0$ (for $t > 0$; also $\gamma > 0$ by assumption), so that $\mathbf{a}_{\text{cor}}$ has the direction of $-\mathbf{b}'$, that is, opposite to the rotation of the Earth. $|\mathbf{a}_{\text{cor}}|$ is maximum at the North Pole and zero at the equator. The projectile B of mass m_0 experiences a force $-m_0 \mathbf{a}_{\text{cor}}$ opposite to $m_0 \mathbf{a}_{\text{cor}}$, which tends to let B deviate from M to the right (and in the Southern Hemisphere, where $\sin \gamma t < 0$, to the left). This deviation has been observed for missiles, rockets, shells, and atmospheric airflow. ■

Curvature and Torsion. *Optional*

This last topic of Sec. 9.5 is optional but completes our discussion of curves relevant to vector calculus.

The **curvature** $\kappa(s)$ of a curve C: $\mathbf{r}(s)$ (s the arc length) at a point P of C measures the rate of change $|\mathbf{u}'(s)|$ of the unit tangent vector $\mathbf{u}(s)$ at P. Hence $\kappa(s)$ measures the deviation of C at P from a straight line (its tangent at P). Since $\mathbf{u}(s) = \mathbf{r}'(s)$, the definition is

$$\kappa(s) = |\mathbf{u}'(s)| = |\mathbf{r}''(s)| \qquad ({}' = d/ds). \tag{22}$$

The **torsion** $\tau(s)$ of C at P measures the rate of change of the **osculating plane** O of curve C at point P. Note that this plane is spanned by $\mathbf{u}$ and $\mathbf{u}'$ and shown in Fig. 212. Hence $\tau(s)$ measures the deviation of C at P from a plane (from O at P). Now the rate of change is also measured by the derivative $\mathbf{b}'$ of a normal vector $\mathbf{b}$ at O. By the definition of vector product, a unit normal vector of O is $\mathbf{b} = \mathbf{u} \times (1/\kappa)\mathbf{u}' = \mathbf{u} \times \mathbf{p}$. Here $\mathbf{p} = (1/\kappa)\mathbf{u}'$ is called the **unit principal normal vector** and $\mathbf{b}$ is called the **unit binormal vector** of C at P. The vectors are labeled in Fig. 212. Here we must assume that $\kappa \neq 0$; hence $\kappa > 0$. The absolute value of the torsion is now defined by

$$|\tau(s)| = |\mathbf{b}'(s)|. \tag{23*}$$

Whereas $\kappa(s)$ is nonnegative, it is practical to give the torsion a sign, motivated by "right-handed" and "left-handed" (see Figs. 204 and 205). This needs a little further calculation. Since $\mathbf{b}$ is a unit vector, it has constant length. Hence $\mathbf{b}'$ is perpendicular

[3]GUSTAVE GASPARD CORIOLIS (1792–1843), French engineer who did research in mechanics.

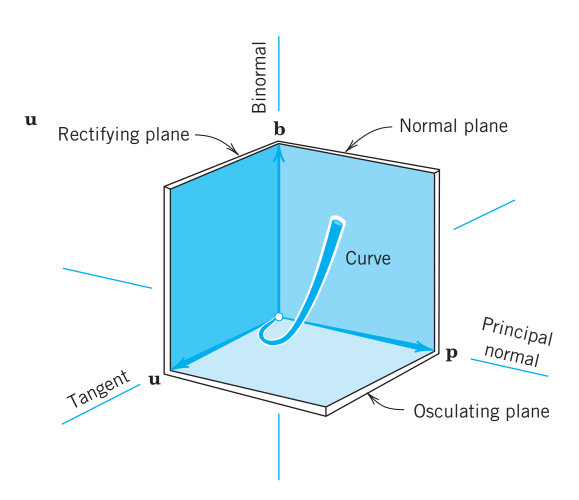


Fig. 212. Trihedron. Unit vectors **u**, **p**, **b** and planes

to **b** (see Example 4 in Sec. 9.4). Now $\mathbf{b}'$ is also perpendicular to **u** because, by the definition of vector product, we have $\mathbf{b} \cdot \mathbf{u} = 0$, $\mathbf{b} \cdot \mathbf{u}' = 0$. This implies

$$(\mathbf{b} \cdot \mathbf{u})' = 0; \qquad \text{that is,} \qquad \mathbf{b}' \cdot \mathbf{u} + \mathbf{b} \cdot \mathbf{u}' = \mathbf{b}' \cdot \mathbf{u} + 0 = 0.$$

Hence if $\mathbf{b}' \neq \mathbf{0}$ at P, it must have the direction of **p** or $-\mathbf{p}$, so that it must be of the form $\mathbf{b}' = -\tau\mathbf{p}$. Taking the dot product of this by **p** and using $\mathbf{p} \cdot \mathbf{p} = 1$ gives

$$\tau(s) = -\mathbf{p}(s) \cdot \mathbf{b}'(s). \tag{23}$$

The minus sign is chosen to make the torsion of a *right-handed* helix *positive* and that of a *left-handed* helix *negative* (Figs. 204 and 205). The orthonormal vector triple **u**, **p**, **b** is called the **trihedron** of C. Figure 212 also shows the names of the three straight lines in the directions of **u**, **p**, **b**, which are the intersections of the **osculating plane**, the **normal plane**, and the **rectifying plane**.

PROBLEM SET 9.5

1–10 PARAMETRIC REPRESENTATIONS

What curves are represented by the following? Sketch them.

1. $[3 + 2\cos t, 2\sin t, 0]$
2. $[a + t, b + 3t, c - 5t]$
3. $[0, t, t^3]$
4. $[-2, 2 + 5\cos t, -1 + 5\sin t]$
5. $[2 + 4\cos t, 1 + \sin t, 0]$
6. $[a + 3\cos \pi t, b - 2\sin \pi t, 0]$
7. $[4\cos t, 4\sin t, 3t]$
8. $[\cosh t, \sinh t, 2]$
9. $[\cos t, \sin 2t, 0]$
10. $[t, 2, 1/t]$

11–20 FIND A PARAMETRIC REPRESENTATION

11. Circle in the plane $z = 1$ with center (3, 2) and passing through the origin.
12. Circle in the yz-plane with center (4, 0) and passing through (0, 3). Sketch it.
13. Straight line through (2, 1, 3) in the direction of $\mathbf{i} + 2\mathbf{j}$.
14. Straight line through (1, 1, 1) and (4, 0, 2). Sketch it.
15. Straight line $y = 4x - 1$, $z = 5x$.
16. The intersection of the circular cylinder of radius 1 about the z-axis and the plane $z = y$.
17. Circle $\frac{1}{2}x^2 + y^2 = 1$, $z = y$.
18. Helix $x^2 + y^2 = 25$, $z = 2\arctan(y/x)$.
19. Hyperbola $4x^2 - 3y^2 = 4$, $z = -2$.

20. Intersection of $2x - y + 3z = 2$ and $x + 2y - z = 3$.

21. Orientation. Explain why setting $t = -t^*$ reverses the orientation of $[a \cos t, a \sin t, 0]$.

22. CAS PROJECT. Curves. Graph the following more complicated curves:

(a) $\mathbf{r}(t) = [2 \cos t + \cos 2t, 2 \sin t - \sin 2t]$ (*Steiner's hypocycloid*).

(b) $\mathbf{r}(t) = [\cos t + k \cos 2t, \sin t - k \sin 2t]$ with $k = 10, 2, 1, \frac{1}{2}, 0, -\frac{1}{2}, -1$.

(c) $\mathbf{r}(t) = [\cos t, \quad \sin 5t]$ (a *Lissajous curve*).

(d) $\mathbf{r}(t) = [\cos t, \quad \sin kt]$. For what k's will it be closed?

(e) $\mathbf{r}(t) = [R \sin \omega t + \omega Rt, \quad R \cos \omega t + R]$ (*cycloid*).

23. CAS PROJECT. Famous Curves in Polar Form. Use your CAS to graph the following curves[4] given in polar form $\rho = \rho(\theta)$, $\rho^2 = x^2 + y^2$, $\tan \theta = y/x$, and investigate their form depending on parameters a and b.

$\rho = a\theta$ *Spiral of Archimedes*

$\rho = ae^{b\theta}$ *Logarithmic spiral*

$\rho = \dfrac{2a \sin^2 \theta}{\cos \theta}$ *Cissoid of Diocles*

$\rho = \dfrac{a}{\cos \theta} + b$ *Conchoid of Nicomedes*

$\rho = a/\theta$ *Hyperbolic spiral*

$\rho = \dfrac{3a \sin 2\theta}{\cos^3 \theta + \sin^3 \theta}$ *Folium of Descartes*

$\rho = 2a \dfrac{\sin 3\theta}{\sin 2\theta}$ *Maclaurin's trisectrix*

$\rho = 2a \cos \theta + b$ *Pascal's snail*

24–28 TANGENT

Given a curve C: $\mathbf{r}(t)$, find a tangent vector $\mathbf{r}'(t)$, a unit tangent vector $\mathbf{u}'(t)$, and the tangent of C at P. Sketch curve and tangent.

24. $\mathbf{r}(t) = [t, \frac{1}{2}t^2, 1]$, P: (2, 2, 1)

25. $\mathbf{r}(t) = [10 \cos t, 1, 10 \sin t]$, P: (6, 1, 8)

26. $\mathbf{r}(t) = [\cos t, \sin t, 9t]$, P: $(1, 0, 18\pi)$

27. $\mathbf{r}(t) = [t, 1/t, 0]$, P: $(2, \frac{1}{2}, 0)$

28. $\mathbf{r}(t) = [t, t^2, t^3]$, P: (1, 1, 1)

29–32 LENGTH

Find the length and sketch the curve.

29. Catenary $\mathbf{r}(t) = [t, \cosh t]$ from $t = 0$ to $t = 1$.

30. Circular helix $\mathbf{r}(t) = [4 \cos t, 4 \sin t, 5t]$ from (4, 0, 0) to $(4, 0, 10\pi)$.

31. Circle $\mathbf{r}(t) = [a \cos t, a \sin t]$ from $(a, 0)$ to $(0, a)$.

32. Hypocycloid $\mathbf{r}(t) = [a \cos^3 t, a \sin^3 t]$, total length.

33. Plane curve. Show that Eq. (10) implies $\ell = \int_a^b \sqrt{1 + y'^2}\, dx$ for the length of a plane curve C: $y = f(x)$, $z = 0$, and $a = x = b$.

34. Polar coordinates $\rho = \sqrt{x^2 + y^2}$, $\theta = \arctan (y/x)$ give

$$\ell = \int_\alpha^\beta \sqrt{\rho^2 + \rho'^2}\, d\theta,$$

where $\rho' = d\rho/d\theta$. Derive this. Use it to find the total length of the **cardioid** $\rho = a(1 - \cos \theta)$. Sketch this curve. *Hint.* Use (10) in App. 3.1.

35–46 CURVES IN MECHANICS

Forces acting on moving objects (cars, airplanes, ships, etc.) require the engineer to know corresponding ***tangential*** and ***normal accelerations***. In Probs. 35–38 find them, along with the ***velocity*** and ***speed***. Sketch the path.

35. Parabola $\mathbf{r}(t) = [t, t^2, 0]$. Find $\mathbf{v}$ and $\mathbf{a}$.

36. Straight line $\mathbf{r}(t) = [8t, 6t, 0]$. Find $\mathbf{v}$ and $\mathbf{a}$.

37. Cycloid $\mathbf{r}(t) = (R \sin \omega t + Rt)\mathbf{i} + (R \cos \omega t + R)\mathbf{j}$. This is the path of a point on the rim of a wheel of radius R that rolls without slipping along the x-axis. Find $\mathbf{v}$ and $\mathbf{a}$ at the maximum y-values of the curve.

38. Ellipse $\mathbf{r} = [\cos t, 2 \sin t, 0]$.

39–42 THE USE OF A CAS

THE USE OF A CAS may greatly facilitate the investigation of more complicated paths, as they occur in gear transmissions and other constructions. To grasp the idea, using a CAS, graph the path and find velocity, speed, and tangential and normal acceleration.

39. $\mathbf{r}(t) = [\cos t + \cos 2t, \sin t - \sin 2t]$

40. $\mathbf{r}(t) = [2 \cos t + \cos 2t, 2 \sin t - \sin 2t]$

41. $\mathbf{r}(t) = [\cos t, \sin 2t, \cos 2t]$

42. $\mathbf{r}(t) = [ct \cos t, ct \sin t, ct]$ $(c \neq 0)$

43. Sun and Earth. Find the acceleration of the Earth toward the sun from (19) and the fact that Earth revolves about the sun in a nearly circular orbit with an almost constant speed of 30 km/s.

44. Earth and moon. Find the centripetal acceleration of the moon toward Earth, assuming that the orbit of the moon is a circle of radius 239,000 miles $= 3.85 \cdot 10^8$ m, and the time for one complete revolution is 27.3 days $= 2.36 \cdot 10^6$ s.

[4] Named after ARCHIMEDES (c. 287–212 B.C.), DESCARTES (Sec. 9.1), DIOCLES (200 B.C.), MACLAURIN (Sec. 15.4), NICOMEDES (250? B.C.) ÉTIENNE PASCAL (1588–1651), father of BLAISE PASCAL (1623–1662).

45. Satellite. Find the speed of an artificial Earth satellite traveling at an altitude of 80 miles above Earth's surface, where $g = 31$ ft/sec^2. (The radius of the Earth is 3960 miles.)

46. Satellite. A satellite moves in a circular orbit 450 miles above Earth's surface and completes 1 revolution in 100 min. Find the acceleration of gravity at the orbit from these data and from the radius of Earth (3960 miles).

47–55 CURVATURE AND TORSION

47. Circle. Show that a circle of radius a has curvature $1/a$.

48. Curvature. Using (22), show that if C is represented by $\mathbf{r}(t)$ with arbitrary t, then

$$(22^*) \quad \kappa(t) = \frac{\sqrt{(\mathbf{r}' \cdot \mathbf{r}')(\mathbf{r}'' \cdot \mathbf{r}'') - (\mathbf{r}' \cdot \mathbf{r}'')^2}}{(\mathbf{r}' \cdot \mathbf{r}')^{3/2}}.$$

49. Plane curve. Using (22*), show that for a curve $y = f(x)$,

$$(22^{**}) \quad \kappa(x) = \frac{|y''|}{(1 + y'^2)^{3/2}} \qquad \left(y' = \frac{dy}{dx}, \text{ etc.}\right).$$

50. Torsion. Using $\mathbf{b} = \mathbf{u} \times \mathbf{p}$ and (23), show that (when $\kappa > 0$)

$$(23^{**}) \quad \tau(s) = (\mathbf{u} \quad \mathbf{p} \quad \mathbf{p}') = (\mathbf{r}' \quad \mathbf{r}'' \quad \mathbf{r}''')/\kappa^2.$$

51. Torsion. Show that if C is represented by $\mathbf{r}(t)$ with arbitrary parameter t, then, assuming $\kappa > 0$ as before,

$$(23^{***}) \quad \tau(t) = \frac{(\mathbf{r}' \quad \mathbf{r}'' \quad \mathbf{r}''')}{(\mathbf{r}' \cdot \mathbf{r}')(\mathbf{r}'' \cdot \mathbf{r}'') - (\mathbf{r}' \cdot \mathbf{r}'')^2}.$$

52. Helix. Show that the helix $[a \cos t, \quad a \sin t, \quad ct]$ can be represented by $[a \cos (s/K), \quad a \sin (s/K), \quad cs/K]$, where $K = \sqrt{a^2 + c^2}$ and s is the arc length. Show that it has constant curvature $\kappa = a/K^2$ and torsion $\tau = c/K^2$.

53. Find the torsion of $C\colon \mathbf{r}(t) = [t, t^2, t^3]$, which looks similar to the curve in Fig. 212.

54. Frenet[5] formulas. Show that

$$\mathbf{u}' = \kappa\,\mathbf{p}, \quad \mathbf{p}' = -\kappa\mathbf{u} + \tau\mathbf{b}, \quad \mathbf{b}' = -\tau\mathbf{p}.$$

55. Obtain κ and τ in Prob. 52 from (22*) and (23***) and the original representation in Prob. 54 with parameter t.

9.6 Calculus Review: Functions of Several Variables. *Optional*

The parametric representations of curves C required vector functions that depended on a *single* variable x, s, or t. We now want to systematically cover vector functions of *several variables*. This optional section is inserted into the book for your convenience and to make the book reasonably self-contained. **Go onto Sec. 9.7 and consult Sec. 9.6 only when needed**. *For partial derivatives*, see App. A3.2.

Chain Rules

Figure 213 shows the notations in the following basic theorem.

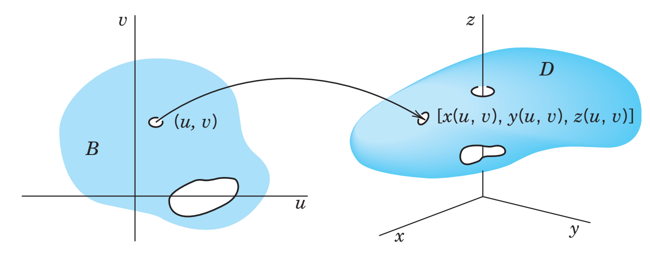


Fig. 213. Notations in Theorem 1

[5]JEAN-FRÉDÉRIC FRENET (1816–1900), French mathematician.

THEOREM 1

Chain Rule

Let $w = f(x, y, z)$ *be continuous and have continuous first partial derivatives in a domain* D *in* xyz*-space. Let* $x = x(u, v)$, $y = y(u, v)$, $z = z(u, v)$ *be functions that are continuous and have first partial derivatives in a domain* B *in the* uv*-plane, where* B *is such that for every point* (u, v) *in* B*, the corresponding point* $[x(u, v), y(u, v), z(u, v)]$ *lies in* D. See Fig. 213. *Then the function*

$$w = f(x(u, v), y(u, v), z(u, v))$$

is defined in B*, has first partial derivatives with respect to* u *and* v *in* B*, and*

$$\textbf{(1)} \qquad \begin{aligned} \frac{\partial w}{\partial u} &= \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial u} \\ \frac{\partial w}{\partial v} &= \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v} + \frac{\partial w}{\partial z}\frac{\partial z}{\partial v}. \end{aligned}$$

In this theorem, a **domain** D is an open connected point set in xyz-space, where "connected" means that any two points of D can be joined by a broken line of finitely many linear segments all of whose points belong to D. "Open" means that every point P of D has a neighborhood (a little ball with center P) all of whose points belong to D. For example, the interior of a cube or of an ellipsoid (the solid without the boundary surface) is a domain.

In calculus, x, y, z are often called the **intermediate variables**, in contrast with the **independent variables** u, v and the **dependent variable** w.

Special Cases of Practical Interest

If $w = f(x, y)$ and $x = x(u, v)$, $y = y(u, v)$ as before, then (1) becomes

$$\textbf{(2)} \qquad \begin{aligned} \frac{\partial w}{\partial u} &= \frac{\partial w}{\partial x}\frac{\partial x}{\partial u} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial u} \\ \frac{\partial w}{\partial v} &= \frac{\partial w}{\partial x}\frac{\partial x}{\partial v} + \frac{\partial w}{\partial y}\frac{\partial y}{\partial v}. \end{aligned}$$

If $w = f(x, y, z)$ and $x = x(t)$, $y = y(t)$, $z = z(t)$, then (1) gives

$$\textbf{(3)} \qquad \frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt} + \frac{\partial w}{\partial z}\frac{dz}{dt}.$$

If $w = f(x, y)$ and $x = x(t)$, $y = y(t)$, then (3) reduces to

(4)
$$\frac{dw}{dt} = \frac{\partial w}{\partial x}\frac{dx}{dt} + \frac{\partial w}{\partial y}\frac{dy}{dt}.$$

Finally, the simplest case $w = f(x)$, $x = x(t)$ gives

(5)
$$\frac{dw}{dt} = \frac{dw}{dx}\frac{dx}{dt}.$$

EXAMPLE 1 **Chain Rule**

If $w = x^2 - y^2$ and we define polar coordinates r, θ by $x = r\cos\theta$, $y = r\sin\theta$, then (2) gives

$$\frac{\partial w}{\partial r} = 2x\cos\theta - 2y\sin\theta = 2r\cos^2\theta - 2r\sin^2\theta = 2r\cos 2\theta$$

$$\frac{\partial w}{\partial \theta} = 2x(-r\sin\theta) - 2y(r\cos\theta) = -2r^2\cos\theta\sin\theta - 2r^2\sin\theta\cos\theta = -2r^2\sin 2\theta.$$ ■

Partial Derivatives on a Surface $z = g(x, y)$

Let $w = f(x, y, z)$ and let $z = g(x, y)$ represent a surface S in space. Then on S the function becomes

$$\tilde{w}(x, y) = f(x, y, g(x, y)).$$

Hence, by (1), the partial derivatives are

(6)
$$\frac{\partial \tilde{w}}{\partial x} = \frac{\partial f}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial x}, \qquad \frac{\partial \tilde{w}}{\partial y} = \frac{\partial f}{\partial y} + \frac{\partial f}{\partial z}\frac{\partial g}{\partial y} \qquad [z = g(x, y)].$$

We shall need this formula in Sec. 10.9.

EXAMPLE 2 **Partial Derivatives on Surface**

Let $w = f = x^3 + y^3 + z^3$ and let $z = g = x^2 + y^2$. Then (6) gives

$$\frac{\partial \tilde{w}}{\partial x} = 3x^2 + 3z^2 \cdot 2x = 3x^2 + 3(x^2 + y^2)^2 \cdot 2x,$$

$$\frac{\partial \tilde{w}}{\partial y} = 3y^2 + 3z^2 \cdot 2y = 3y^2 + 3(x^2 + y^2)^2 \cdot 2y.$$

We confirm this by substitution, using $w(x, y) = x^3 + y^3 + (x^2 + y^2)^3$, that is,

$$\frac{\partial \tilde{w}}{\partial x} = 3x^2 + 3(x^2 + y^2)^2 \cdot 2x, \qquad \frac{\partial \tilde{w}}{\partial y} = 3y^2 + 3(x^2 + y^2)^2 \cdot 2y.$$ ■

Mean Value Theorems

THEOREM 2

Mean Value Theorem

Let $f(x, y, z)$ be continuous and have continuous first partial derivatives in a domain D in xyz-space. Let P_0: (x_0, y_0, z_0) and P: $(x_0 + h, y_0 + k, z_0 + l)$ be points in D such that the straight line segment P_0P joining these points lies entirely in D. Then

$$f(x_0 + h, y_0 + k, z_0 + l) - f(x_0, y_0, z_0) = h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y} + l\frac{\partial f}{\partial z}, \tag{7}$$

the partial derivatives being evaluated at a suitable point of that segment.

Special Cases

For a function $f(x, y)$ of two variables (satisfying assumptions as in the theorem), formula (7) reduces to (Fig. 214)

$$f(x_0 + h, y_0 + k) - f(x_0, y_0) = h\frac{\partial f}{\partial x} + k\frac{\partial f}{\partial y}, \tag{8}$$

and, for a function $f(x)$ of a single variable, (7) becomes

$$f(x_0 + h) - f(x_0) = h\frac{\partial f}{\partial x}, \tag{9}$$

where in (9), the domain D is a segment of the x-axis and the derivative is taken at a suitable point between x_0 and $x_0 + h$.

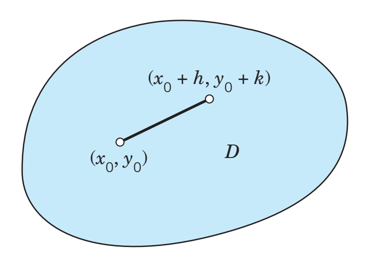


Fig. 214. Mean value theorem for a function of two variables [Formula (8)]

9.7 Gradient of a Scalar Field. Directional Derivative

We shall see that *some* of the vector fields that occur in applications—not all of them!—can be obtained from scalar fields. Using scalar fields instead of vector fields is of a considerable advantage because scalar fields are easier to use than vector fields. It is the

"gradient" that allows us to obtain vector fields from scalar fields, and thus the gradient is of great practical importance to the engineer.

DEFINITION 1

Gradient

The setting is that we are given a scalar function $f(x, y, z)$ that is defined and differentiable in a domain in 3-space with Cartesian coordinates x, y, z. We denote the **gradient** of that function by grad f or ∇f (read **nabla** f). Then the qradient of $f(x, y, z)$ is defined as the vector function

(1) $$\operatorname{grad} f = \nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}.$$

Remarks. For a definition of the gradient in curvilinear coordinates, see App. 3.4. As a quick example, if $f(x, y, z) = 2y^3 + 4xz + 3x$, then grad $f = [4z + 3, 6y^2, 4x]$. Furthermore, we will show later in this section that (1) actually does define a vector.

The notation ∇f is suggested by the *differential operator* ∇ (read *nabla*) defined by

(1*) $$\nabla = \frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}.$$

Gradients are useful in several ways, notably in giving the rate of change of $f(x, y, z)$ in any direction in space, in obtaining surface normal vectors, and in deriving vector fields from scalar fields, as we are going to show in this section.

Directional Derivative

From calculus we know that the partial derivatives in (1) give the rates of change of $f(x, y, z)$ in the directions of the three coordinate axes. It seems natural to extend this and ask for the rate of change of f *in an arbitrary direction* in space. This leads to the following concept.

DEFINITION 2

Directional Derivative

The directional derivative $D_{\mathbf{b}}f$ or df/ds of a function $f(x, y, z)$ at a point P in the direction of a vector $\mathbf{b}$ is defined by (see Fig. 215)

(2) $$D_{\mathbf{b}}f = \frac{df}{ds} = \lim_{s \to 0} \frac{f(Q) - f(P)}{s}.$$

Here Q is a variable point on the straight line L in the direction of $\mathbf{b}$, and $|s|$ is the distance between P and Q. Also, $s > 0$ if Q lies in the direction of $\mathbf{b}$ (as in Fig. 215), $s < 0$ if Q lies in the direction of $-\mathbf{b}$, and $s = 0$ if $Q = P$.

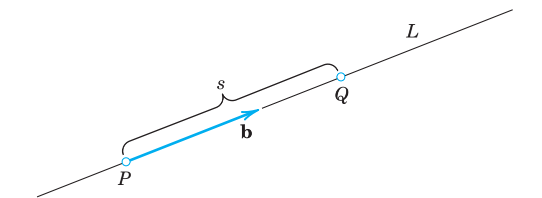


Fig. 215. Directional derivative

The next idea is to use Cartesian xyz-coordinates and for **b** a unit vector. Then the line L is given by

$$\mathbf{r}(s) = x(s)\mathbf{i} + y(s)\mathbf{j} + z(s)\mathbf{k} = \mathbf{p}_0 + s\mathbf{b} \qquad (|\mathbf{b}| = 1) \tag{3}$$

where $\mathbf{p}_0$ the position vector of P. Equation (2) now shows that $D_{\mathbf{b}}f = df/ds$ is the derivative of the function $f(x(s), y(s), z(s))$ with respect to the arc length s of L. Hence, assuming that f has continuous partial derivatives and applying the chain rule [formula (3) in the previous section], we obtain

$$D_{\mathbf{b}}f = \frac{df}{ds} = \frac{\partial f}{\partial x}x' + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial z}z' \tag{4}$$

where primes denote derivatives with respect to s (which are taken at $s = 0$). But here, differentiating (3) gives $\mathbf{r}' = x'\mathbf{i} + y'\mathbf{j} + z'\mathbf{k} = \mathbf{b}$. Hence (4) is simply the inner product of grad f and **b** [see (2), Sec. 9.2]; that is,

$$D_{\mathbf{b}}f = \frac{df}{ds} = \mathbf{b} \bullet \operatorname{grad} f \qquad (|\mathbf{b}| = 1). \tag{5}$$

ATTENTION! If the direction is given by a vector **a** of any length ($\neq 0$), then

$$D_{\mathbf{a}}f = \frac{df}{ds} = \frac{1}{|\mathbf{a}|}\mathbf{a} \bullet \operatorname{grad} f. \tag{5*}$$

EXAMPLE 1 **Gradient. Directional Derivative**

Find the directional derivative of $f(x, y, z) = 2x^2 + 3y^2 + z^2$ at P: (2, 1, 3) in the direction of $\mathbf{a} = [1, 0, -2]$.

Solution. grad $f = [4x, 6y, 2z]$ gives at P the vector grad $f(P) = [8, 6, 6]$. From this and (5*) we obtain, since $|\mathbf{a}| = \sqrt{5}$,

$$D_{\mathbf{a}}f(P) = \frac{1}{\sqrt{5}}[1, 0, -2] \bullet [8, 6, 6] = \frac{1}{\sqrt{5}}(8 + 0 - 12) = -\frac{4}{\sqrt{5}} = -1.789.$$

The minus sign indicates that at P the function f is decreasing in the direction of **a**. ■

Gradient Is a Vector. Maximum Increase

Here is a finer point of mathematics that concerns the consistency of our theory: grad f in (1) *looks* like a vector—after all, it has three components! But to prove that it *actually is* a vector, since it is defined in terms of components depending on the Cartesian coordinates, we must show that grad f has a length and direction independent of the choice of those coordinates. See proof of Theorem 1. In contrast, $[\partial f/\partial x, 2\partial f/\partial y, \partial f/\partial z]$ also looks like a vector but does not have a length and direction independent of the choice of Cartesian coordinates.

Incidentally, the direction makes the gradient eminently useful: grad f *points in the direction of maximum increase of* f.

THEOREM 1

Use of Gradient: Direction of Maximum Increase

Let $f(P) = f(x, y, z)$ *be a scalar function having continuous first partial derivatives in some domain* B *in space. Then* grad f *exists in* B *and is a vector, that is, its length and direction are independent of the particular choice of Cartesian coordinates. If* grad $f(P) \neq \mathbf{0}$ *at some point* P*, it has the direction of maximum increase of* f *at* P.

PROOF From (5) and the definition of inner product [(1) in Sec. 9.2] we have

$$D_{\mathbf{b}} f = |\mathbf{b}||\text{grad } f| \cos \gamma = |\text{grad } f| \cos \gamma \tag{6}$$

where γ is the angle between $\mathbf{b}$ and grad f. Now f is a scalar function. Hence its value at a point P depends on P but not on the particular choice of coordinates. The same holds for the arc length s of the line L in Fig. 215, hence also for $D_{\mathbf{b}} f$. Now (6) shows that $D_{\mathbf{b}} f$ is maximum when $\cos \gamma = 1$, $\gamma = 0$, and then $D_{\mathbf{b}} f = |\text{grad } f|$. It follows that the length and direction of grad f are independent of the choice of coordinates. Since $\gamma = 0$ if and only if $\mathbf{b}$ has the direction of grad f, the latter is the direction of maximum increase of f at P, provided grad $f \neq \mathbf{0}$ at P. Make sure that you understood the proof to get a good feel for mathematics.

Gradient as Surface Normal Vector

Gradients have an important application in connection with surfaces, namely, as surface normal vectors, as follows. Let S be a surface represented by $f(x, y, z) = c = \text{const}$, where f is differentiable. Such a surface is called a **level surface** of f, and for different c we get different level surfaces. Now let C be a curve on S through a point P of S. As a curve in space, C has a representation $\mathbf{r}(t) = [x(t), y(t), z(t)]$. For C to lie on the surface S, the components of $\mathbf{r}(t)$ must satisfy $f(x, y, z) = c$, that is,

$$f(x(t), y(t), z(t)) = c. \tag{7}$$

Now a tangent vector of C is $\mathbf{r}'(t) = [x'(t), y'(t), z'(t)]$. And the tangent vectors of all curves on S passing through P will generally form a plane, called the **tangent plane** of S at P. (Exceptions occur at edges or cusps of S, for instance, at the apex of the cone in Fig. 217.) The normal of this plane (the straight line through P perpendicular to the tangent plane) is called the **surface normal** to S at P. A vector in the direction of the surface

normal is called a **surface normal vector** of S at P. We can obtain such a vector quite simply by differentiating (7) with respect to t. By the chain rule,

$$\frac{\partial f}{\partial x}x' + \frac{\partial f}{\partial y}y' + \frac{\partial f}{\partial z}z' = (\operatorname{grad} f) \cdot \mathbf{r}' = 0.$$

Hence grad f is orthogonal to all the vectors $\mathbf{r}'$ in the tangent plane, so that it is a normal vector of S at P. Our result is as follows (see Fig. 216).

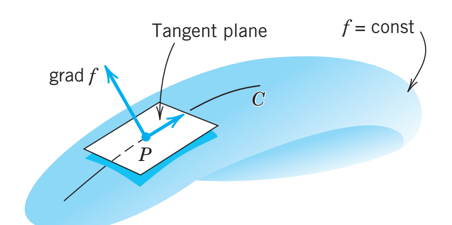


Fig. 216. Gradient as surface normal vector

THEOREM 2

Gradient as Surface Normal Vector

Let f be a differentiable scalar function in space. Let $f(x, y, z) = c = const$ represent a surface S. Then if the gradient of f at a point P of S is not the zero vector, it is a normal vector of S at P.

EXAMPLE 2

Gradient as Surface Normal Vector. Cone

Find a unit normal vector $\mathbf{n}$ of the cone of revolution $z^2 = 4(x^2 + y^2)$ at the point P: (1, 0, 2).

Solution. The cone is the level surface $f = 0$ of $f(x, y, z) = 4(x^2 + y^2) - z^2$. Thus (Fig. 217)

$$\operatorname{grad} f = [8x, \quad 8y, \quad -2z], \qquad \operatorname{grad} f(P) = [8, \quad 0, \quad -4]$$

$$\mathbf{n} = \frac{1}{|\operatorname{grad} f(P)|}\operatorname{grad} f(P) = \left[\frac{2}{\sqrt{5}}, \quad 0, \quad -\frac{1}{\sqrt{5}}\right].$$

$\mathbf{n}$ points downward since it has a negative z-component. The other unit normal vector of the cone at P is $-\mathbf{n}$. ■

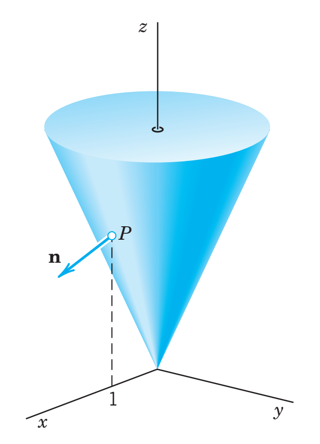


Fig. 217. Cone and unit normal vector **n**

Vector Fields That Are Gradients of Scalar Fields ("Potentials")

At the beginning of this section we mentioned that some vector fields have the advantage that they can be obtained from scalar fields, which can be worked with more easily. Such a vector field is given by a vector function $\mathbf{v}(P)$, which is obtained as the gradient of a scalar function, say, $\mathbf{v}(P) = \text{grad}\, f(P)$. The function $f(P)$ is called a *potential function* or a **potential** of $\mathbf{v}(P)$. Such a $\mathbf{v}(P)$ and the corresponding vector field are called **conservative** because in such a vector field, energy is conserved; that is, no energy is lost (or gained) in displacing a body (or a charge in the case of an electrical field) from a point P to another point in the field and back to P. We show this in Sec. 10.2.

Conservative fields play a central role in physics and engineering. A basic application concerns the gravitational force (see Example 3 in Sec. 9.4) and we show that it has a potential which satisfies Laplace's equation, the most important partial differential equation in physics and its applications.

THEOREM 3

Gravitational Field. Laplace's Equation

The force of attraction

$$\mathbf{p} = -\frac{c}{r^3}\mathbf{r} = -c\left[\frac{x - x_0}{r^3}, \frac{y - y_0}{r^3}, \frac{z - z_0}{r^3}\right] \tag{8}$$

between two particles at points P_0: (x_0, y_0, z_0) and P: (x, y, z) (as given by Newton's law of gravitation) *has the potential* $f(x, y, z) = c/r$, *where* $r\ (> 0)$ *is the distance between* P_0 *and* P.

Thus $\mathbf{p} = \text{grad}\, f = \text{grad}\,(c/r)$. *This potential* f *is a solution of* **Laplace's equation**

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0. \tag{9}$$

[$\nabla^2 f$ (read *nabla squared* f) is called the **Laplacian** of f.]

PROOF That distance is $r = ((x - x_0)^2 + (y - y_0)^2 + (z - z_2)^2)^{1/2}$. The key observation now is that for the components of $\mathbf{p} = [p_1, p_2, p_3]$ we obtain by partial differentiation

$$\frac{\partial}{\partial x}\left(\frac{1}{r}\right) = \frac{-2(x - x_0)}{2[(x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2]^{3/2}} = -\frac{x - x_0}{r^3} \tag{10a}$$

and similarly

$$\begin{aligned} \frac{\partial}{\partial y}\left(\frac{1}{r}\right) &= -\frac{y - y_0}{r^3}, \\ \frac{\partial}{\partial z}\left(\frac{1}{r}\right) &= -\frac{z - z_0}{r^3}. \end{aligned} \tag{10b}$$

From this we see that, indeed, **p** is the gradient of the scalar function $f = c/r$. The second statement of the theorem follows by partially differentiating (10), that is,

$$\frac{\partial^2}{\partial x^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(x - x_0)^2}{r^5},$$

$$\frac{\partial^2}{\partial y^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(y - y_0)^2}{r^5},$$

$$\frac{\partial^2}{\partial z^2}\left(\frac{1}{r}\right) = -\frac{1}{r^3} + \frac{3(z - z_0)^2}{r^5},$$

and then adding these three expressions. Their common denominator is r^5. Hence the three terms $-1/r^3$ contribute $-3r^2$ to the numerator, and the three other terms give the sum

$$3(x - x_0)^2 + 3(y - y_0)^2 + 3(z - z_0)^2 = 3r^2,$$

so that the numerator is 0, and we obtain (9). ■

$\nabla^2 f$ is also denoted by Δf. The differential operator

(11) $$\nabla^2 = \Delta = \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}$$

(read "nabla squared" or "delta") is called the **Laplace operator**. It can be shown that the field of force produced by any distribution of masses is given by a vector function that is the gradient of a scalar function f, and f satisfies (9) in any region that is free of matter.

The great importance of the Laplace equation also results from the fact that there are other laws in physics that are of the same form as Newton's law of gravitation. For instance, in electrostatics the force of attraction (or repulsion) between two particles of opposite (or like) charge Q_1 and Q_2 is

(12) $$\mathbf{p} = \frac{k}{r^3}\mathbf{r}$$ (**Coulomb's law**[6]).

Laplace's equation will be discussed in detail in Chaps. 12 and 18.

A method for finding out whether a given vector field has a potential will be explained in Sec. 9.9.

[6]CHARLES AUGUSTIN DE COULOMB (1736–1806), French physicist and engineer. Coulomb's law was derived by him from his own very precise measurements.

PROBLEM SET 9.7

1–6 CALCULATION OF GRADIENTS

Find grad f. Graph some level curves $f = \text{const}$. Indicate ∇f by arrows at some points of these curves.

1. $f = (x + 1)(2y - 1)$

2. $f = 9x^2 + 4y^2$

3. $f = y/x$

4. $(y + 6)^2 + (x - 4)^2$

5. $f = x^4 + y^4$

6. $f = (x^2 - y^2)/(x^2 + y^2)$

7–10 USEFUL FORMULAS FOR GRADIENT AND LAPLACIAN

Prove and illustrate by an example.

7. $\nabla(f^n) = nf^{n-1}\nabla f$

8. $\nabla(fg) = f\nabla g + g\nabla f$

9. $\nabla(f/g) = (1/g^2)(g\nabla f - f\nabla g)$

10. $\nabla^2(fg) = g\nabla^2 f + 2\nabla f \cdot \nabla g + f\nabla^2 g$

11–15 USE OF GRADIENTS. ELECTRIC FORCE

The force in an electrostatic field given by $f(x, y, z)$ has the direction of the gradient. Find ∇f and its value at P.

11. $f = xy$, $P: (-4, 5)$

12. $f = x/(x^2 + y^2)$, $P: (1, 1)$

13. $f = \ln(x^2 + y^2)$, $P: (8, 6)$

14. $f = (x^2 + y^2 + z^2)^{-1/2}$ $P: (12, 0, 16)$

15. $f = 4x^2 + 9y^2 + z^2$, $P: (5, -1, -11)$

16. For what points $P: (x, y, z)$ does ∇f with $f = 25x^2 + 9y^2 + 16z^2$ have the direction from P to the origin?

17. Same question as in Prob. 16 when $f = 25x^2 + 4y^2$.

18–23 VELOCITY FIELDS

Given the velocity potential f of a flow, find the velocity $\mathbf{v} = \nabla f$ of the field and its value $\mathbf{v}(P)$ at P. Sketch $\mathbf{v}(P)$ and the curve $f = \text{const}$ passing through P.

18. $f = x^2 - 6x - y^2$, $P: (-1, 5)$

19. $f = \cos x \cosh y$, $P: (\frac{1}{2}\pi, \ln 2)$

20. $f = x(1 + (x^2 + y^2)^{-1})$, $P: (1, 1)$

21. $f = e^x \cos y$, $P: (1, \frac{1}{2}\pi)$

22. At what points is the flow in Prob. 21 directed vertically upward?

23. At what points is the flow in Prob. 21 horizontal?

24–27 HEAT FLOW

Experiments show that in a temperature field, heat flows in the direction of maximum decrease of temperature T. Find this direction in general and at the given point P. Sketch that direction at P as an arrow.

24. $T = 3x^2 - 2y^2$, $P: (2.5, 1.8)$

25. $T = z/(x^2 + y^2)$, $P: (0, 1, 2)$

26. $T = x^2 + y^2 + 4z^2$, $P: (2, -1, 2)$

27. CAS PROJECT. Isotherms. Graph some curves of constant temperature ("isotherms") and indicate directions of heat flow by arrows when the temperature equals **(a)** $x^3 - 3xy^2$, **(b)** $\sin x \sinh y$, and **(c)** $e^x \cos y$.

28. Steepest ascent. If $z(x, y) = 3000 - x^2 - 9y^2$ [meters] gives the elevation of a mountain at sea level, what is the direction of steepest ascent at $P: (4, 1)$?

29. Gradient. What does it mean if $|\nabla f(P)| > |\nabla f(Q)|$ at two points P and Q in a scalar field?

9.8 Divergence of a Vector Field

Vector calculus owes much of its importance in engineering and physics to the gradient, divergence, and curl. From a scalar field we can obtain a vector field by the gradient (Sec. 9.7). Conversely, from a vector field we can obtain a scalar field by the divergence or another vector field by the curl (to be discussed in Sec. 9.9). These concepts were suggested by basic physical applications. This will be evident from our examples.

To begin, let $\mathbf{v}(x, y, z)$ be a differentiable vector function, where x, y, z are Cartesian coordinates, and let v_1, v_2, v_3 be the components of $\mathbf{v}$. Then the function

$$\text{div } \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z} \tag{1}$$

is called the **divergence** of $\mathbf{v}$ or the *divergence of the vector field defined by* $\mathbf{v}$. For example, if

$$\mathbf{v} = [3xz, 2xy, -yz^2] = 3xz\mathbf{i} + 2xy\mathbf{j} - yz^2\mathbf{k}, \qquad \text{then} \qquad \operatorname{div} \mathbf{v} = 3z + 2x - 2yz.$$

Another common notation for the divergence is

$$\begin{aligned} \operatorname{div} \mathbf{v} = \nabla \bullet \mathbf{v} &= \left[\frac{\partial}{\partial x}, \frac{\partial}{\partial \mathbf{y}}, \frac{\partial}{\partial \mathbf{z}}\right] \bullet [v_1, v_2, v_3] \\ &= \left(\frac{\partial}{\partial x}\mathbf{i} + \frac{\partial}{\partial y}\mathbf{j} + \frac{\partial}{\partial z}\mathbf{k}\right) \bullet (v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}) \\ &= \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}, \end{aligned}$$

with the understanding that the "product" $(\partial/\partial x)v_1$ in the dot product means the partial derivative $\partial v_1/\partial x$, etc. This is a convenient notation, but nothing more. Note that $\nabla \bullet \mathbf{v}$ means the scalar div $\mathbf{v}$, whereas ∇f means the vector grad f defined in Sec. 9.7.

In Example 2 we shall see that the divergence has an important physical meaning. Clearly, the values of a function that characterizes a physical or geometric property must be independent of the particular choice of coordinates. In other words, these values must be invariant with respect to coordinate transformations. Accordingly, the following theorem should hold.

THEOREM 1

Invariance of the Divergence

The divergence div $\mathbf{v}$ *is a scalar function, that is, its values depend only on the points in space (and, of course, on* $\mathbf{v}$*) but not on the choice of the coordinates in* (1)*, so that with respect to other Cartesian coordinates* x^*, y^*, z^* *and corresponding components* v_1^*, v_2^*, v_3^* *of* $\mathbf{v}$,

$$\operatorname{div} \mathbf{v} = \frac{\partial v_1^*}{\partial x^*} + \frac{\partial v_2^*}{\partial y^*} + \frac{\partial v_3^*}{\partial z^*}. \tag{2}$$

We shall prove this theorem in Sec. 10.7, using integrals.

Presently, let us turn to the more immediate practical task of gaining a feel for the significance of the divergence. Let $f(x, y, z)$ be a twice differentiable scalar function. Then its gradient exists,

$$\mathbf{v} = \operatorname{grad} f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$

and we can differentiate once more, the first component with respect to x, the second with respect to y, the third with respect to z, and then form the divergence,

$$\operatorname{div} \mathbf{v} = \operatorname{div} (\operatorname{grad} f) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2}.$$

Hence we have the basic result that *the divergence of the gradient is the Laplacian* (Sec. 9.7),

$$\text{div}\,(\text{grad}\, f) = \nabla^2 f. \tag{3}$$

EXAMPLE 1 Gravitational Force. Laplace's Equation

The gravitational force **p** in Theorem 3 of the last section is the gradient of the scalar function $f(x, y, z) = c/r$, which satisfies Laplaces equation $\nabla^2 f = 0$. According to (3) this implies that div $\mathbf{p} = 0$ $(r > 0)$. ■

The following example from hydrodynamics shows the physical significance of the divergence of a vector field. We shall get back to this topic in Sec. 10.8 and add further physical details.

EXAMPLE 2 Flow of a Compressible Fluid. Physical Meaning of the Divergence

We consider the motion of a fluid in a region R having no **sources** or **sinks** in R, that is, no points at which fluid is produced or disappears. The concept of **fluid state** is meant to cover also gases and vapors. Fluids in the restricted sense, or liquids, such as water or oil, have very small compressibility, which can be neglected in many problems. In contrast, gases and vapors have high compressibility. Their density ρ (= mass per unit volume) depends on the coordinates x, y, z in space and may also depend on time t. We assume that our fluid is compressible. We consider the flow through a rectangular box B of small edges Δx, Δy, Δz parallel to the coordinate axes as shown in Fig. 218. (Here Δ is a standard notation for small quantities and, of course, has nothing to do with the notation for the Laplacian in (11) of Sec. 9.7.) The box B has the volume $\Delta V = \Delta x\, \Delta y\, \Delta z$. Let $\mathbf{v} = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ be the velocity vector of the motion. We set

$$\mathbf{u} = \rho\mathbf{v} = [u_1, u_2, u_3] = u_1\mathbf{i} + u_2\mathbf{j} + u_3\mathbf{k} \tag{4}$$

and assume that **u** and **v** are continuously differentiable vector functions of x, y, z, and t, that is, they have first partial derivatives which are continuous. Let us calculate the change in the mass included in B by considering the **flux** across the boundary, that is, the total loss of mass leaving B per unit time. Consider the flow through the left of the three faces of B that are visible in Fig. 218, whose area is $\Delta x\,\Delta z$. Since the vectors $v_1\mathbf{i}$ and $v_3\,\mathbf{k}$ are parallel to that face, the components v_1 and v_3 of **v** contribute nothing to this flow. Hence the mass of fluid entering through that face during a short time interval Δt is given approximately by

$$(\rho v_2)_y\, \Delta x\, \Delta z\, \Delta t = (u_2)_y\, \Delta x\, \Delta z\, \Delta t,$$

where the subscript y indicates that this expression refers to the left face. The mass of fluid leaving the box B through the opposite face during the same time interval is approximately $(u_2)_{y+\Delta y}\, \Delta x\, \Delta z\, \Delta t$, where the subscript $y + \Delta y$ indicates that this expression refers to the right face (which is not visible in Fig. 218). The difference

$$\Delta u_2\, \Delta x\, \Delta z\, \Delta t = \frac{\Delta u_2}{\Delta y}\, \Delta V\, \Delta t \qquad [\Delta u_2 = (u_2)_{y+\Delta y} - (u_2)_y]$$

is the approximate loss of mass. Two similar expressions are obtained by considering the other two pairs of parallel faces of B. If we add these three expressions, we find that the total loss of mass in B during the time interval Δt is approximately

$$\left(\frac{\Delta u_1}{\Delta x} + \frac{\Delta u_2}{\Delta y} + \frac{\Delta u_3}{\Delta z}\right) \Delta V\, \Delta t,$$

where

$$\Delta u_1 = (u_1)_{x+\Delta x} - (u_1)_x \qquad \text{and} \qquad \Delta u_3 = (u_3)_{z+\Delta z} - (u_3)_z.$$

This loss of mass in B is caused by the time rate of change of the density and is thus equal to

$$-\frac{\partial \rho}{\partial t}\,\Delta V\,\Delta t.$$

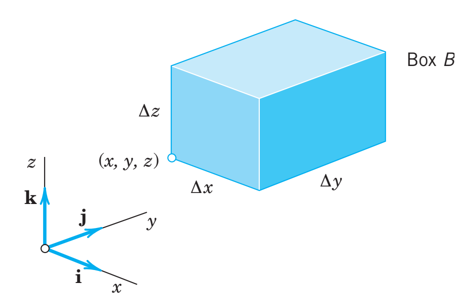


Fig. 218. Physical interpretation of the divergence

If we equate both expressions, divide the resulting equation by $\Delta V \Delta t$, and let Δx, Δy, Δz, and Δt approach zero, then we obtain

$$\operatorname{div} \mathbf{u} = \operatorname{div} (\rho \mathbf{v}) = -\frac{\partial \rho}{\partial t}$$

or

$$\frac{\partial \rho}{\partial t} + \operatorname{div} (\rho \mathbf{v}) = 0. \tag{5}$$

This important relation is called the *condition for the conservation of mass* or the **continuity equation** *of a compressible fluid flow.*

If the flow is **steady**, that is, independent of time, then $\partial \rho / \partial t = 0$ and the continuity equation is

$$\operatorname{div} (\rho \mathbf{v}) = 0. \tag{6}$$

If the density ρ is constant, so that the fluid is incompressible, then equation (6) becomes

$$\operatorname{div} \mathbf{v} = 0. \tag{7}$$

This relation is known as the **condition of incompressibility**. It expresses the fact that the balance of outflow and inflow for a given volume element is zero at any time. Clearly, the assumption that the flow has no sources or sinks in R is essential to our argument. $\mathbf{v}$ is also referred to as **solenoidal**.

From this discussion you should conclude and remember that, roughly speaking, ***the divergence measures outflow minus inflow***. ■

Comment. The **divergence theorem** of Gauss, an integral theorem involving the divergence, follows in the next chapter (Sec. 10.7).

PROBLEM SET 9.8

1–6 CALCULATION OF THE DIVERGENCE

Find div $\mathbf{v}$ and its value at P.

1. $\mathbf{v} = [x^2, 4y^2, 9z^2]$, P: $(-1, 0, \frac{1}{2})$

2. $\mathbf{v} = [0, \cos xyz, \sin xyz]$, P: $(2, \frac{1}{2}\pi, 0)$

3. $\mathbf{v} = (x^2 + y^2)^{-1}[x, y]$

4. $\mathbf{v} = [v_1(y, z), v_2(z, x), v_3(x, y)]$, P: $(3, 1, -1)$]

5. $\mathbf{v} = x^2y^2z^2[x, y, z]$, P: $(3, -1, 4)$

6. $\mathbf{v} = (x^2 + y^2 + z^2)^{-3/2}[x, y, z]$

7. For what v_3 is $\mathbf{v} = [e^x \cos y, e^x \sin y, v_3]$ solenoidal?

8. Let $\mathbf{v} = [x, y, v_3]$. Find a v_3 such that **(a)** div $\mathbf{v} > 0$ everywhere, **(b)** div $\mathbf{v} > 0$ if $|z| < 1$ and div $\mathbf{v} < 0$ if $|z| > 1$.

9. PROJECT. Useful Formulas for the Divergence. Prove

(a) $\operatorname{div}(k\mathbf{v}) = k \operatorname{div} \mathbf{v}$ (k constant)

(b) $\operatorname{div}(f\mathbf{v}) = f \operatorname{div} \mathbf{v} + \mathbf{v} \bullet \nabla f$

(c) $\operatorname{div}(f\nabla g) = f\nabla^2 g + \nabla f \bullet \nabla g$

(d) $\operatorname{div}(f\nabla g) - \operatorname{div}(g\nabla f) = f\nabla^2 g - g\nabla^2 f$

Verify (b) for $f = e^{xyz}$ and $\mathbf{v} = ax\mathbf{i} + by\mathbf{j} + cz\mathbf{k}$. Obtain the answer to Prob. 6 from (b). Verify (c) for $f = x^2 - y^2$ and $g = e^{x+y}$. Give examples of your own for which (a)–(d) are advantageous.

10. CAS EXPERIMENT. Visualizing the Divergence. Graph the given velocity field $\mathbf{v}$ of a fluid flow in a square centered at the origin with sides parallel to the coordinate axes. Recall that the divergence measures outflow minus inflow. By looking at the flow near the sides of the square, can you see whether div $\mathbf{v}$ must be positive or negative or may perhaps be zero? Then calculate div $\mathbf{v}$. First do the given flows and then do some of your own. Enjoy it.

(a) $\mathbf{v} = \mathbf{i}$

(b) $\mathbf{v} = x\mathbf{i}$

(c) $\mathbf{v} = x\mathbf{i} - y\mathbf{j}$

(d) $\mathbf{v} = x\mathbf{i} + y\mathbf{j}$

(e) $\mathbf{v} = -x\mathbf{i} - y\mathbf{j}$

(f) $\mathbf{v} = (x^2 + y^2)^{-1}(-y\mathbf{i} + x\mathbf{j})$

11. Incompressible flow. Show that the flow with velocity vector $\mathbf{v} = y\mathbf{i}$ is incompressible. Show that the particles that at time $t = 0$ are in the cube whose faces are portions of the planes $x = 0, x = 1, y = 0, y = 1, z = 0, z = 1$ occupy at $t = 1$ the volume 1.

12. Compressible flow. Consider the flow with velocity vector $\mathbf{v} = x\mathbf{i}$. Show that the individual particles have the position vectors $\mathbf{r}(t) = c_1 e^t \mathbf{i} + c_2\mathbf{j} + c_3\mathbf{k}$ with constant c_1, c_2, c_3. Show that the particles that at $t = 0$ are in the cube of Prob. 11 at $t = 1$ occupy the volume e.

13. Rotational flow. The velocity vector $\mathbf{v}(x, y, z)$ of an incompressible fluid rotating in a cylindrical vessel is of the form $\mathbf{v} = \mathbf{w} \times \mathbf{r}$, where $\mathbf{w}$ is the (constant) rotation vector; see Example 5 in Sec. 9.3. Show that div $\mathbf{v} = 0$. Is this plausible because of our present Example 2?

14. Does $\operatorname{div} \mathbf{u} = \operatorname{div} \mathbf{v}$ imply $\mathbf{u} = \mathbf{v}$ or $\mathbf{u} = \mathbf{v} + \mathbf{k}$ ($\mathbf{k}$ constant)? Give reason.

15–20 LAPLACIAN

Calculate $\nabla^2 f$ by Eq. (3). Check by direct differentiation. Indicate when (3) is simpler. Show the details of your work.

15. $f = \cos^2 x + \sin^2 y$

16. $f = e^{xyz}$

17. $f = \ln(x^2 + y^2)$

18. $f = z - \sqrt{x^2 + y^2}$

19. $f = 1/(x^2 + y^2 + z^2)$

20. $f = e^{2x} \cosh 2y$

9.9 Curl of a Vector Field

The concepts of gradient (Sec. 9.7), divergence (Sec. 9.8), and curl are of fundamental importance in vector calculus and frequently applied in vector fields. In this section we define and discuss the concept of the curl and apply it to several engineering problems.

Let $\mathbf{v}(x, y, z) = [v_1, v_2, v_3] = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k}$ be a differentiable vector function of the Cartesian coordinates x, y, z. Then the **curl** *of the vector function* $\mathbf{v}$ or *of the vector field given by* $\mathbf{v}$ is defined by the "symbolic" determinant

$$\text{(1)} \qquad \operatorname{curl} \mathbf{v} = \nabla \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix}$$

$$= \left(\frac{\partial v_3}{\partial y} - \frac{\partial v_2}{\partial z}\right)\mathbf{i} + \left(\frac{\partial v_1}{\partial z} - \frac{\partial v_3}{\partial x}\right)\mathbf{j} + \left(\frac{\partial v_2}{\partial x} - \frac{\partial v_1}{\partial y}\right)\mathbf{k}.$$

This is the formula when x, y, z are *right-handed.* If they are *left-handed*, the determinant has a minus sign in front (just as in (2**) in Sec. 9.3).

Instead of curl $\mathbf{v}$ one also uses the notation rot $\mathbf{v}$. This is suggested by "rotation," an application explored in Example 2. Note that curl $\mathbf{v}$ is a vector, as shown in Theorem 3.

EXAMPLE 1 Curl of a Vector Function

Let $\mathbf{v} = [yz, \quad 3zx, \quad z] = yz\mathbf{i} + 3zx\mathbf{j} + z\mathbf{k}$ with right-handed x, y, z. Then (1) gives

$$\text{curl } \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ yz & 3zx & z \end{vmatrix} = -3x\mathbf{i} + y\mathbf{j} + (3z - z)\mathbf{k} = -3x\mathbf{i} + y\mathbf{j} + 2z\mathbf{k}.$$ ■

The curl has many applications. A typical example follows. More about the nature and significance of the curl will be considered in Sec. 10.9.

EXAMPLE 2 Rotation of a Rigid Body. Relation to the Curl

We have seen in Example 5, Sec. 9.3, that a rotation of a rigid body B about a fixed axis in space can be described by a vector $\mathbf{w}$ of magnitude ω in the direction of the axis of rotation, where ω (>0) is the angular speed of the rotation, and $\mathbf{w}$ is directed so that the rotation appears clockwise if we look in the direction of $\mathbf{w}$. According to (9), Sec. 9.3, the velocity field of the rotation can be represented in the form

$$\mathbf{v} = \mathbf{w} \times \mathbf{r}$$

where $\mathbf{r}$ is the position vector of a moving point with respect to a Cartesian coordinate system *having the origin on the axis of rotation*. Let us choose right-handed Cartesian coordinates such that the axis of rotation is the z-axis. Then (see Example 2 in Sec. 9.4)

$$\mathbf{w} = [0, \quad 0, \quad \omega] = \omega\mathbf{k}, \qquad \mathbf{v} = \mathbf{w} \times \mathbf{r} = [-\omega y, \quad \omega x, \quad 0] = -\omega y\mathbf{i} + \omega x\mathbf{j}.$$

Hence

$$\text{curl } \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ -\omega y & \omega x & 0 \end{vmatrix} = [0, \quad 0, \quad 2\omega] = 2\omega\mathbf{k} = 2\mathbf{w}.$$

This proves the following theorem. ■

THEOREM 1 Rotating Body and Curl

The curl of the velocity field of a rotating rigid body has the direction of the axis of the rotation, and its magnitude equals twice the angular speed of the rotation.

Next we show how the grad, div, and curl are interrelated, thereby shedding further light on the nature of the curl.

THEOREM 2

Grad, Div, Curl

Gradient fields are ***irrotational***. *That is, if a continuously differentiable vector function is the gradient of a scalar function f, then its curl is the zero vector,*

$$\text{curl}\,(\text{grad}\, f) = \mathbf{0}. \tag{2}$$

Furthermore, the divergence of the curl of a twice continuously differentiable vector function **v** *is zero,*

$$\text{div}\,(\text{curl}\ \mathbf{v}) = 0. \tag{3}$$

PROOF Both (2) and (3) follow directly from the definitions by straightforward calculation. In the proof of (3) the six terms cancel in pairs. ■

EXAMPLE 3 Rotational and Irrotational Fields

The field in Example 2 is not irrotational. A similar velocity field is obtained by stirring tea or coffee in a cup. The gravitational field in Theorem 3 of Sec. 9.7 has curl $\mathbf{p} = \mathbf{0}$. It is an irrotational gradient field. ■

The term "irrotational" for curl $\mathbf{v} = \mathbf{0}$ is suggested by the use of the curl for characterizing the rotation in a field. If a gradient field occurs elsewhere, not as a velocity field, it is usually called **conservative** (see Sec. 9.7). Relation (3) is plausible because of the interpretation of the curl as a rotation and of the divergence as a flux (see Example 2 in Sec. 9.8).

Finally, since the curl is defined in terms of coordinates, we should do what we did for the gradient in Sec. 9.7, namely, to find out whether the curl is a vector. This is true, as follows.

THEOREM 3

Invariance of the Curl

curl **v** is a vector. It has a length and a direction that are independent of the particular choice of a Cartesian coordinate system in space.

PROOF The proof is quite involved and shown in App. 4.

We have completed our discussion of vector differential calculus. The companion Chap. 10 on vector integral calculus follows and makes use of many concepts covered in this chapter, including dot and cross products, parametric representation of curves C, along with grad, div, and curl.

PROBLEM SET 9.9

1. **WRITING REPORT. Grad, div, curl.** List the definitions and most important facts and formulas for grad, div, curl, and ∇^2. Use your list to write a corresponding report of 3–4 pages, with examples of your own. No proofs.
2. **(a)** What direction does curl **v** have if **v** is parallel to the yz-plane? **(b)** If, moreover, **v** is independent of x?
3. Prove Theorem 2. Give two examples for (2) and (3) each.

4–8 CALCULUTION OF CURL

Find curl **v** for **v** given with respect to right-handed Cartesian coordinates. Show the details of your work.

4. $\mathbf{v} = [2y^2, 5x, 0]$
5. $\mathbf{v} = xyz[x, y, z]$
6. $\mathbf{v} = (x^2 + y^2 + z^2)^{-3/2}[x, y, z]$
7. $\mathbf{v} = [0, 0, e^{-x} \sin y]$
8. $\mathbf{v} = [e^{-z^2}, e^{-x^2}, e^{-y^2}]$

9–13 FLUID FLOW

Let **v** be the velocity vector of a steady fluid flow. Is the flow irrotational? Incompressible? Find the streamlines (the paths of the particles). *Hint.* See the answers to Probs. 9 and 11 for a determination of a path.

9. $\mathbf{v} = [0, 3z^2, 0]$
10. $\mathbf{v} = [\sec x, \csc x, 0]$
11. $\mathbf{v} = [y, -2x, 0]$
12. $\mathbf{v} = [-y, x, \pi]$
13. $\mathbf{v} = [x, y, -z]$

14. **PROJECT. Useful Formulas for the Curl.** Assuming sufficient differentiability, show that
   (a) $\operatorname{curl}(\mathbf{u} + \mathbf{v}) = \operatorname{curl} \mathbf{u} + \operatorname{curl} \mathbf{v}$
   (b) $\operatorname{div}(\operatorname{curl} \mathbf{v}) = 0$
   (c) $\operatorname{curl}(f\mathbf{v}) = (\operatorname{grad} f) \times \mathbf{v} + f \operatorname{curl} \mathbf{v}$
   (d) $\operatorname{curl}(\operatorname{grad} f) = \mathbf{0}$
   (e) $\operatorname{div}(\mathbf{u} \times \mathbf{v}) = \mathbf{v} \cdot \operatorname{curl} \mathbf{u} - \mathbf{u} \cdot \operatorname{curl} \mathbf{v}$

15–20 DIV AND CURL

With respect to right-handed coordinates, let $\mathbf{u} = [y, z, x]$, $\mathbf{v} = [yz, zx, xy]$, $f = xyz$, and $g = x + y + z$. Find the given expressions. Check your result by a formula in Proj. 14 if applicable.

15. $\operatorname{curl}(\mathbf{u} + \mathbf{v})$, $\operatorname{curl} \mathbf{v}$
16. $\operatorname{curl}(g\mathbf{v})$
17. $\mathbf{v} \cdot \operatorname{curl} \mathbf{u}$, $\mathbf{u} \cdot \operatorname{curl} \mathbf{v}$, $\mathbf{u} \cdot \operatorname{curl} \mathbf{u}$
18. $\operatorname{div}(\mathbf{u} \times \mathbf{v})$
19. $\operatorname{curl}(g\mathbf{u} + \mathbf{v})$, $\operatorname{curl}(g\mathbf{u})$
20. $\operatorname{div}(\operatorname{grad}(fg))$

CHAPTER 9 REVIEW QUESTIONS AND PROBLEMS

1. What is a vector? A vector function? A vector field? A scalar? A scalar function? A scalar field? Give examples.
2. What is an inner product, a vector product, a scalar triple product? What applications motivate these products?
3. What are right-handed and left-handed coordinates? When is this distinction important?
4. When is a vector product the zero vector? What is orthogonality?
5. How is the derivative of a vector function defined? What is its significance in geometry and mechanics?
6. If $\mathbf{r}(t)$ represents a motion, what are $\mathbf{r}'(t)$, $|\mathbf{r}'(t)|$, $\mathbf{r}''(t)$, and $|\mathbf{r}''(t)|$?
7. Can a moving body have constant speed but variable velocity? Nonzero acceleration?
8. What do you know about directional derivatives? Their relation to the gradient?
9. Write down the definitions and explain the significance of grad, div, and curl.
10. Granted sufficient differentiability, which of the following expressions make sense? $f \operatorname{curl} \mathbf{v}$, $\mathbf{v} \operatorname{curl} f$, $\mathbf{u} \times \mathbf{v}$, $\mathbf{u} \times \mathbf{v} \times \mathbf{w}$, $f \cdot \mathbf{v}$, $f \cdot (\mathbf{v} \times \mathbf{w})$, $\mathbf{u} \cdot (\mathbf{v} \times \mathbf{w})$, $\mathbf{v} \times \operatorname{curl} \mathbf{v}$, $\operatorname{div}(f\mathbf{v})$, $\operatorname{curl}(f\mathbf{v})$, and $\operatorname{curl}(f \cdot \mathbf{v})$.

11–19 ALGEBRAIC OPERATIONS FOR VECTORS

Let $\mathbf{a} = [4, 7, 0]$, $\mathbf{b} = [3, -1, 5]$, $\mathbf{c} = [-6, 2, 0]$, and $\mathbf{d} = [1, -2, 8]$. Calculate the following expressions. Try to make a sketch.

11. $\mathbf{a} \cdot \mathbf{c}$, $3\mathbf{b} \cdot 8\mathbf{d}$, $24\mathbf{d} \cdot \mathbf{b}$, $\mathbf{a} \cdot \mathbf{a}$
12. $\mathbf{a} \times \mathbf{c}$, $\mathbf{b} \times \mathbf{d}$, $\mathbf{d} \times \mathbf{b}$, $\mathbf{a} \times \mathbf{a}$
13. $\mathbf{b} \times \mathbf{c}$, $\mathbf{c} \times \mathbf{b}$, $\mathbf{c} \times \mathbf{c}$, $\mathbf{c} \cdot \mathbf{c}$
14. $5(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}$, $\mathbf{a} \cdot (5\mathbf{b} \times \mathbf{c})$, $(5\mathbf{a}\ \mathbf{b}\ \mathbf{c})$, $5(\mathbf{a} \cdot \mathbf{b}) \times \mathbf{c}$
15. $6(\mathbf{a} \times \mathbf{b}) \times \mathbf{d}$, $\mathbf{a} \times 6(\mathbf{b} \times \mathbf{d})$, $2\mathbf{a} \times 3\mathbf{b} \times \mathbf{d}$
16. $(1/|\mathbf{a}|)\mathbf{a}$, $(1/|\mathbf{b}|)\mathbf{b}$, $\mathbf{a} \cdot \mathbf{b}/|\mathbf{b}|$, $\mathbf{a} \cdot \mathbf{b}/|\mathbf{a}|$
17. $(\mathbf{a}\ \mathbf{b}\ \mathbf{d})$, $(\mathbf{b}\ \mathbf{a}\ \mathbf{d})$, $(\mathbf{b}\ \mathbf{d}\ \mathbf{a})$
18. $|\mathbf{a} + \mathbf{b}|$, $|\mathbf{a}| + |\mathbf{b}|$
19. $\mathbf{a} \times \mathbf{b} - \mathbf{b} \times \mathbf{a}$, $(\mathbf{a} \times \mathbf{c}) \cdot \mathbf{c}$, $|\mathbf{a} \times \mathbf{b}|$
20. **Commutativity.** When is $\mathbf{u} \times \mathbf{v} = \mathbf{v} \times \mathbf{u}$? When is $\mathbf{u} \cdot \mathbf{v} = \mathbf{v} \cdot \mathbf{u}$?
21. **Resultant, equilibrium.** Find **u** such that **u** and **a**, **b**, **c**, **d** above and **u** are in equilibrium.
22. **Resultant.** Find the most general **v** such that the resultant of **v**, **a**, **b**, **c** (see above) is parallel to the *yz*-plane.
23. **Angle.** Find the angle between **a** and **c**. Between **b** and **d**. Sketch **a** and **c**.
24. **Planes.** Find the angle between the two planes P_1: $4x - y + 3z = 12$ and P_2: $x + 2y + 4z = 4$. Make a sketch.
25. **Work.** Find the work done by $\mathbf{q} = [5, 2, 0]$ in the displacement from (1, 1, 0) to (4, 3, 0).
26. **Component.** When is the component of a vector **v** in the direction of a vector **w** equal to the component of **w** in the direction of **v**?
27. **Component.** Find the component of $\mathbf{v} = [4, 7, 0]$ in the direction of $\mathbf{w} = [2, 2, 0]$. Sketch it.

28. Moment. When is the moment of a force equal to zero?

29. Moment. A force $\mathbf{p} = [4, 2, 0]$ is acting in a line through (2, 3, 0). Find its moment vector about the center (5, 1, 0) of a wheel.

30. Velocity, acceleration. Find the velocity, speed, and acceleration of the motion given by $\mathbf{r}(t) = [3 \cos t, 3 \sin t, 4t]$ (t = time) at the point P: $(3/\sqrt{2}, 3/\sqrt{2}, \pi)$.

31. Tetrahedron. Find the volume if the vertices are (0, 0, 0), (3, 1, 2), (2, 4, 0), (5, 4, 0).

32–40 GRAD, DIV, CURL, ∇^2, $D_\mathbf{v}f$

Let $f = xy - yz$, $\mathbf{v} = [2y, 2z, 4x + z]$, and $\mathbf{w} = [3z^2, x^2 - y^2, y^2]$. Find:

32. grad f and f grad f at P: (2, 7, 0)

33. div $\mathbf{v}$, div $\mathbf{w}$

34. curl $\mathbf{v}$, curl $\mathbf{w}$

35. div (grad f), $\nabla^2 f$, $\nabla^2(xyf)$

36. (curl $\mathbf{w}$) • $\mathbf{v}$ at (4, 0, 2)

37. grad (div $\mathbf{w}$)

38. $D_\mathbf{v} f$ at P: (1, 1, 2)

39. $D_\mathbf{w} f$ at P: (3, 0, 2)

40. $\mathbf{v} \bullet ((\text{curl } \mathbf{w}) \times \mathbf{v})$

SUMMARY OF CHAPTER 9

Vector Differential Calculus. Grad, Div, Curl

All vectors of the form $\mathbf{a} = [a_1, a_2, a_3] = a_1\mathbf{i} + a_2\mathbf{j} + a_3\mathbf{k}$ constitute the **real vector space** R^3 with componentwise vector addition

$$[a_1, a_2, a_3] + [b_1, b_2, b_3] = [a_1 + b_1, a_2 + b_2, a_3 + b_3] \tag{1}$$

and componentwise scalar multiplication (c a scalar, a real number)

$$c[a_1, a_2, a_3] = [ca_1, ca_2, ca_3] \qquad \text{(Sec. 9.1).} \tag{2}$$

For instance, the *resultant* of forces $\mathbf{a}$ and $\mathbf{b}$ is the sum $\mathbf{a} + \mathbf{b}$.

The **inner product** or **dot product** of two vectors is defined by

$$\mathbf{a} \bullet \mathbf{b} = |\mathbf{a}||\mathbf{b}| \cos \gamma = a_1 b_1 + a_2 b_2 + a_3 b_3 \qquad \text{(Sec. 9.2)} \tag{3}$$

where γ is the angle between $\mathbf{a}$ and $\mathbf{b}$. This gives for the **norm** or **length** $|\mathbf{a}|$ of $\mathbf{a}$

$$|\mathbf{a}| = \sqrt{\mathbf{a} \bullet \mathbf{a}} = \sqrt{a_1^2 + a_2^2 + a_3^2} \tag{4}$$

as well as a formula for γ. If $\mathbf{a} \bullet \mathbf{b} = 0$, we call $\mathbf{a}$ and $\mathbf{b}$ **orthogonal**. The dot product is suggested by the *work* $W = \mathbf{p} \bullet \mathbf{d}$ done by a force $\mathbf{p}$ in a displacement $\mathbf{d}$.

The **vector product** or **cross product** $\mathbf{v} = \mathbf{a} \times \mathbf{b}$ is a vector of length

$$|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin \gamma \qquad \text{(Sec. 9.3)} \tag{5}$$

and perpendicular to both $\mathbf{a}$ and $\mathbf{b}$ such that $\mathbf{a}$, $\mathbf{b}$, $\mathbf{v}$ form a *right-handed* triple. In terms of components with respect to right-handed coordinates,

$$\mathbf{a} \times \mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix} \qquad \text{(Sec. 9.3).} \tag{6}$$

The vector product is suggested, for instance, by moments of forces or by rotations. CAUTION! This multiplication is *anti*commutative, $\mathbf{a} \times \mathbf{b} = -\mathbf{b} \times \mathbf{a}$, and is *not* associative.

An (oblique) box with edges $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ has volume equal to the absolute value of the **scalar triple product**

$$(7) \qquad (\mathbf{a} \quad \mathbf{b} \quad \mathbf{c}) = \mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) = (\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}.$$

Sections 9.4–9.9 extend differential calculus to vector functions

$$\mathbf{v}(t) = [v_1(t), v_2(t), v_3(t)] = v_1(t)\mathbf{i} + v_2(t)\mathbf{j} + v_3(t)\mathbf{k}$$

and to vector functions of more than one variable (see below). The derivative of $\mathbf{v}(t)$ is

$$(8) \qquad \mathbf{v}' = \frac{d\mathbf{v}}{dt} = \lim_{\Delta t \to 0} \frac{\mathbf{v}(t + \Delta t) - \mathbf{v}(t)}{\Delta t} = [v_1', v_2', v_3'] = v_1'\mathbf{i} + v_2'\mathbf{j} + v_3'\mathbf{k}.$$

Differentiation rules are as in calculus. They imply (Sec. 9.4)

$$(\mathbf{u} \cdot \mathbf{v})' = \mathbf{u}' \cdot \mathbf{v} + \mathbf{u} \cdot \mathbf{v}', \qquad (\mathbf{u} \times \mathbf{v})' = \mathbf{u}' \times \mathbf{v} + \mathbf{u} \times \mathbf{v}'.$$

Curves C in space represented by the position vector $\mathbf{r}(t)$ have $\mathbf{r}'(t)$ as a **tangent vector** (the **velocity** in mechanics when t is time), $\mathbf{r}'(s)$ (s arc length, Sec. 9.5) as the *unit tangent vector,* and $|\mathbf{r}''(s)| = \kappa$ as the *curvature* (the *acceleration* in mechanics).

Vector functions $\mathbf{v}(x, y, z) = [v_1(x, y, z), v_2(x, y, z), v_3(x, y, z)]$ represent vector fields in space. Partial derivatives with respect to the Cartesian coordinates x, y, z are obtained componentwise, for instance,

$$\frac{\partial \mathbf{v}}{\partial x} = \left[\frac{\partial v_1}{\partial x}, \frac{\partial v_2}{\partial x}, \frac{\partial v_3}{\partial x}\right] = \frac{\partial v_1}{\partial x}\mathbf{i} + \frac{\partial v_2}{\partial x}\mathbf{j} + \frac{\partial v_3}{\partial x}\mathbf{k} \qquad \text{(Sec. 9.6).}$$

The **gradient** of a scalar function f is

$$(9) \qquad \operatorname{grad} f = \nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right] \qquad \text{(Sec. 9.7).}$$

The **directional derivative** of f in the direction of a vector $\mathbf{a}$ is

$$(10) \qquad D_{\mathbf{a}} f = \frac{df}{ds} = \frac{1}{|\mathbf{a}|}\mathbf{a} \cdot \nabla f \qquad \text{(Sec. 9.7).}$$

The **divergence** of a vector function $\mathbf{v}$ is

$$(11) \qquad \operatorname{div} \mathbf{v} = \nabla \cdot \mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}. \qquad \text{(Sec. 9.8).}$$

The **curl** of **v** is

$$\text{curl}\,\mathbf{v} = \nabla \times \mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix} \tag{12}$$ (Sec. 9.9)

or minus the determinant if the coordinates are left-handed.

Some basic formulas for grad, div, curl are (Secs. 9.7–9.9)

$$\begin{aligned} \nabla(fg) &= f\nabla g + g\nabla f \\ \nabla(f/g) &= (1/g^2)(g\nabla f - f\nabla g) \end{aligned} \tag{13}$$

$$\begin{aligned} \text{div}\,(f\mathbf{v}) &= f\,\text{div}\,\mathbf{v} + \mathbf{v} \bullet \nabla f \\ \text{div}\,(f\nabla g) &= f\nabla^2 g + \nabla f \bullet \nabla g \end{aligned} \tag{14}$$

$$\begin{aligned} \nabla^2 f &= \text{div}\,(\nabla f) \\ \nabla^2(fg) &= g\nabla^2 f + 2\nabla f \bullet \nabla g + f\nabla^2 g \end{aligned} \tag{15}$$

$$\begin{aligned} \text{curl}\,(f\mathbf{v}) &= \nabla f \times \mathbf{v} + f\,\text{curl}\,\mathbf{v} \\ \text{div}\,(\mathbf{u} \times \mathbf{v}) &= \mathbf{v} \bullet \text{curl}\,\mathbf{u} - \mathbf{u} \bullet \text{curl}\,\mathbf{v} \end{aligned} \tag{16}$$

$$\begin{aligned} \text{curl}\,(\nabla f) &= \mathbf{0} \\ \text{div}\,(\text{curl}\,\mathbf{v}) &= 0. \end{aligned} \tag{17}$$

For grad, div, curl, and ∇^2 in **curvilinear coordinates** see App. A3.4.

CHAPTER 10

Vector Integral Calculus. Integral Theorems

Vector integral calculus can be seen as a generalization of regular integral calculus. You may wish to review integration. (To refresh your memory, there is an optional review section on double integrals; see Sec. 10.3.)

Indeed, vector integral calculus extends integrals as known from regular calculus to integrals over curves, called *line integrals* (Secs. 10.1, 10.2), surfaces, called *surface integrals* (Sec. 10.6), and solids, called *triple integrals* (Sec. 10.7). The beauty of vector integral calculus is that we can transform these different integrals into one another. You do this to simplify evaluations, that is, one type of integral might be easier to solve than another, such as in potential theory (Sec. 10.8). More specifically, Green's theorem in the plane allows you to transform line integrals into double integrals, or conversely, double integrals into line integrals, as shown in Sec. 10.4. Gauss's convergence theorem (Sec. 10.7) converts surface integrals into triple integrals, and vice-versa, and Stokes's theorem deals with converting line integrals into surface integrals, and vice-versa.

This chapter is a companion to Chapter 9 on vector differential calculus. From Chapter 9, you will need to know inner product, curl, and divergence and how to parameterize curves. The root of the transformation of the integrals was largely physical intuition. Since the corresponding formulas involve the divergence and the curl, the study of this material will lead to a deeper physical understanding of these two operations.

Vector integral calculus is very important to the engineer and physicist and has many applications in solid mechanics, in fluid flow, in heat problems, and others.

Prerequisite: Elementary integral calculus, Secs. 9.7–9.9
Sections that may be omitted in a shorter course: 10.3, 10.5, 10.8
References and Answers to Problems: App. 1 Part B, App. 2

10.1 Line Integrals

The concept of a line integral is a simple and natural generalization of a definite integral

$$\int_a^b f(x)\,dx. \tag{1}$$

Recall that, in (1), we integrate the function $f(x)$, also known as the integrand, from $x = a$ along the x-axis to $x = b$. Now, in a line integral, we shall integrate a given function, also

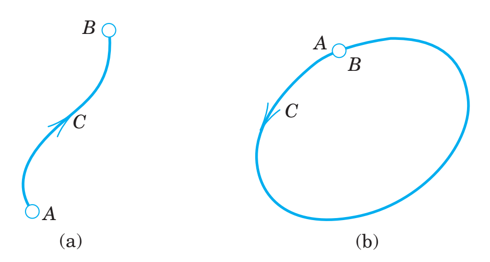


Fig. 219. Oriented curve

called the **integrand**, along a curve C in space or in the plane. (Hence curve integral would be a better name but line integral is standard).

This requires that we represent the curve C by a parametric representation (as in Sec. 9.5)

(2) $$\mathbf{r}(t) = [x(t), y(t), z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k} \qquad (a \leqq t \leqq b).$$

The curve C is called the **path of integration**. Look at Fig. 219a. The path of integration goes from A to B. Thus A: $\mathbf{r}(a)$ is its initial point and B: $\mathbf{r}(b)$ is its terminal point. C is now *oriented*. The direction from A to B, in which t increases is called the positive direction on C. We mark it by an arrow. The points A and B may coincide, as it happens in Fig. 219b. Then C is called a **closed path**.

C is called a **smooth curve** if it has at each point a unique tangent whose direction varies continuously as we move along C. We note that $\mathbf{r}(t)$ in (2) is differentiable. Its derivative $\mathbf{r}'(t) = d\mathbf{r}/dt$ is continuous and different from the zero vector at every point of C.

General Assumption

In this book, every path of integration of a line integral is assumed to be **piecewise smooth**, *that is, it consists of* ***finitely many*** *smooth curves.*

For example, the boundary curve of a square is piecewise smooth. It consists of four smooth curves or, in this case, line segments which are the four sides of the square.

Definition and Evaluation of Line Integrals

A **line integral** of a vector function $\mathbf{F}(\mathbf{r})$ over a curve C: $\mathbf{r}(t)$ is defined by

(3) $$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t)\, dt \qquad \mathbf{r}' = \frac{d\mathbf{r}}{dt}$$

where $\mathbf{r}(t)$ is the parametric representation of C as given in (2). (The dot product was defined in Sec. 9.2.) Writing (3) in terms of components, with $d\mathbf{r} = [dx, \; dy, \; dz]$ as in Sec. 9.5 and $' = d/dt$, we get

(3′) $$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\, dx + F_2\, dy + F_3\, dz) = \int_a^b (F_1 x' + F_2 y' + F_3 z')\, dt.$$

If the path of integration C in (3) is a *closed* curve, then instead of

$$\int_C \qquad \text{we also write} \qquad \oint_C .$$

Note that the integrand in (3) is a scalar, not a vector, because we take the dot product. Indeed, $\mathbf{F} \cdot \mathbf{r}'/|\mathbf{r}'|$ is the tangential component of $\mathbf{F}$. (For "component" see (11) in Sec. 9.2.)

We see that the integral in (3) on the right is a definite integral of a function of t taken over the interval $a \leqq t \leqq b$ on the t-axis in the ***positive*** direction: The direction of increasing t. This definite integral exists for continuous $\mathbf{F}$ and piecewise smooth C, because this makes $\mathbf{F} \cdot \mathbf{r}'$ piecewise continuous.

Line integrals (3) arise naturally in mechanics, where they give the work done by a force $\mathbf{F}$ in a displacement along C. This will be explained in detail below. We may thus call the line integral (3) the **work integral**. Other forms of the line integral will be discussed later in this section.

EXAMPLE 1 Evaluation of a Line Integral in the Plane

Find the value of the line integral (3) when $\mathbf{F}(\mathbf{r}) = [-y, -xy] = -y\mathbf{i} - xy\mathbf{j}$ and C is the circular arc in Fig. 220 from A to B.

Solution. We may represent C by $\mathbf{r}(t) = [\cos t, \sin t] = \cos t\,\mathbf{i} + \sin t\,\mathbf{j}$, where $0 \leqq t \leqq \pi/2$. Then $x(t) = \cos t$, $y(t) = \sin t$, and

$$\mathbf{F}(\mathbf{r}(t)) = -y(t)\mathbf{i} - x(t)y(t)\mathbf{j} = [-\sin t, \quad -\cos t \sin t] = -\sin t\,\mathbf{i} - \cos t \sin t\,\mathbf{j}.$$

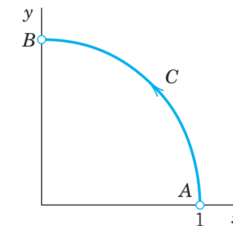


Fig. 220. Example 1

By differentiation, $\mathbf{r}'(t) = [-\sin t, \quad \cos t] = -\sin t\,\mathbf{i} + \cos t\,\mathbf{j}$, so that by (3) [use (10) in App. 3.1; set $\cos t = u$ in the second term]

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_0^{\pi/2} [-\sin t, -\cos t \sin t] \cdot [-\sin t, \cos t]\, dt = \int_0^{\pi/2} (\sin^2 t - \cos^2 t \sin t)\, dt$$

$$= \int_0^{\pi/2} \frac{1}{2}(1 - \cos 2t)\, dt - \int_1^0 u^2(-du) = \frac{\pi}{4} - 0 - \frac{1}{3} \approx 0.4521.$$

EXAMPLE 2 Line Integral in Space

The evaluation of line integrals in space is practically the same as it is in the plane. To see this, find the value of (3) when $\mathbf{F}(\mathbf{r}) = [z, x, y] = z\mathbf{i} + x\mathbf{j} + y\mathbf{k}$ and C is the helix (Fig. 221)

$$\mathbf{r}(t) = [\cos t, \sin t, 3t] = \cos t\,\mathbf{i} + \sin t\,\mathbf{j} + 3t\mathbf{k} \qquad (0 \leqq t \leqq 2\pi). \tag{4}$$

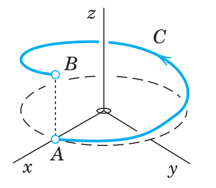
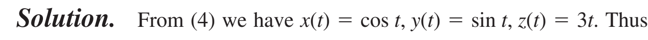


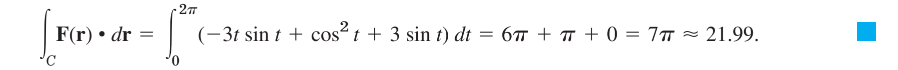
Fig. 221. Example 2



Solution. From (4) we have $x(t) = \cos t$, $y(t) = \sin t$, $z(t) = 3t$. Thus

$$\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = (3t\,\mathbf{i} + \cos t\,\mathbf{j} + \sin t\,\mathbf{k}) \cdot (-\sin t\,\mathbf{i} + \cos t\,\mathbf{j} + 3\mathbf{k}).$$

The dot product is $3t(-\sin t) + \cos^2 t + 3 \sin t$. Hence (3) gives



$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_0^{2\pi} (-3t \sin t + \cos^2 t + 3 \sin t)\, dt = 6\pi + \pi + 0 = 7\pi \approx 21.99.$$

Simple general properties of the line integral (3) follow directly from corresponding properties of the definite integral in calculus, namely,

$$\int_C k\mathbf{F} \cdot d\mathbf{r} = k \int_C \mathbf{F} \cdot d\mathbf{r} \qquad (k \text{ constant}) \tag{5a}$$

$$\int_C (\mathbf{F} + \mathbf{G}) \cdot d\mathbf{r} = \int_C \mathbf{F} \cdot d\mathbf{r} + \int_C \mathbf{G} \cdot d\mathbf{r} \tag{5b}$$

$$\int_C \mathbf{F} \cdot d\mathbf{r} = \int_{C_1} \mathbf{F} \cdot d\mathbf{r} + \int_{C_2} \mathbf{F} \cdot d\mathbf{r} \qquad \text{(Fig. 222)} \tag{5c}$$

Fig. 222. Formula (5c)

where in (5c) the path C is subdivided into two arcs C_1 and C_2 that have the same orientation as C (Fig. 222). In (5b) the orientation of C is the same in all three integrals. If the sense of integration along C is reversed, the value of the integral is multiplied by -1. However, we note the following independence if the sense is preserved.

THEOREM 1

Direction-Preserving Parametric Transformations

Any representations of C that give the same positive direction on C also yield the same value of the line integral (3).

PROOF The proof follows by the chain rule. Let $\mathbf{r}(t)$ be the given representation with $a \leqq t \leqq b$ as in (3). Consider the transformation $t = \phi(t^*)$ which transforms the t interval to $a^* \leqq t^* \leqq b^*$ and has a positive derivative dt/dt^*. We write $\mathbf{r}(t) = \mathbf{r}(\phi(t^*)) = \mathbf{r}^*(t^*)$. Then $dt = (dt/dt^*)\, dt^*$ and

$$\begin{aligned} \int_C \mathbf{F}(\mathbf{r}^*) \cdot d\mathbf{r}^* &= \int_{a^*}^{b^*} \mathbf{F}(\mathbf{r}(\phi(t^*))) \cdot \frac{d\mathbf{r}}{dt}\frac{dt}{dt^*}\, dt^* \\ &= \int_a^b \mathbf{F}(\mathbf{r}(t)) \cdot \frac{d\mathbf{r}}{dt}\, dt = \int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}. \end{aligned}$$

■

Motivation of the Line Integral (3): Work Done by a Force

The work W done by a *constant* force $\mathbf{F}$ in the displacement along a *straight* segment $\mathbf{d}$ is $W = \mathbf{F} \cdot \mathbf{d}$; see Example 2 in Sec. 9.2. This suggests that we define the work W done by a *variable* force $\mathbf{F}$ in the displacement along a curve C: $\mathbf{r}(t)$ as the limit of sums of works done in displacements along small chords of C. We show that this definition amounts to defining W by the line integral (3).

For this we choose points $t_0\,(=a) < t_1 < \cdots < t_n\,(=b)$. Then the work ΔW_m done by $\mathbf{F}(\mathbf{r}(t_m))$ in the straight displacement from $\mathbf{r}(t_m)$ to $\mathbf{r}(t_{m+1})$ is

$$\Delta W_m = \mathbf{F}(\mathbf{r}(t_m)) \cdot [\mathbf{r}(t_{m+1}) - \mathbf{r}(t_m)] \approx \mathbf{F}(\mathbf{r}(t_m)) \cdot \mathbf{r}'(t_m)\Delta t_m \qquad (\Delta t_m = \Delta t_{m+1} - t_m).$$

The sum of these n works is $W_n = \Delta W_0 + \cdots + \Delta W_{n-1}$. If we choose points and consider W_n for every n arbitrarily but so that the greatest Δt_m approaches zero as $n \to \infty$, then the limit of W_n as $n \to \infty$ is the line integral (3). This integral exists because of our general assumption that $\mathbf{F}$ is continuous and C is piecewise smooth; this makes $\mathbf{r}'(t)$ continuous, except at finitely many points where C may have corners or cusps. ■

EXAMPLE 3 Work Done by a Variable Force

If **F** in Example 1 is a force, the work done by **F** in the displacement along the quarter-circle is 0.4521, measured in suitable units, say, newton-meters (nt · m, also called joules, abbreviation J; see also inside front cover). Similarly in Example 2. ■

EXAMPLE 4 Work Done Equals the Gain in Kinetic Energy

Let **F** be a force, so that (3) is work. Let t be time, so that $d\mathbf{r}/dt = \mathbf{v}$, velocity. Then we can write (3) as

$$W = \int_C \mathbf{F} \bullet d\mathbf{r} = \int_a^b \mathbf{F}(\mathbf{r}(t)) \bullet \mathbf{v}(t)\, dt. \tag{6}$$

Now by Newton's second law, that is, force = mass × acceleration, we get

$$\mathbf{F} = m\mathbf{r}''(t) = m\mathbf{v}'(t),$$

where m is the mass of the body displaced. Substitution into (5) gives [see (11), Sec. 9.4]

$$W = \int_a^b m\mathbf{v}' \bullet \mathbf{v}\, dt = \int_a^b m\left(\frac{\mathbf{v} \bullet \mathbf{v}}{2}\right)' dt = \frac{m}{2}|\mathbf{v}|^2 \Big|_{t=a}^{t=b}.$$

On the right, $m|\mathbf{v}|^2/2$ is the kinetic energy. Hence *the work done equals the gain in kinetic energy*. This is a basic law in mechanics. ■

Other Forms of Line Integrals

The line integrals

$$\int_C F_1\, dx, \qquad \int_C F_2\, dy, \qquad \int_C F_3\, dz \tag{7}$$

are special cases of (3) when $\mathbf{F} = F_1\mathbf{i}$ or $F_2\mathbf{j}$ or $F_3\mathbf{k}$, respectively.

Furthermore, without taking a dot product as in (3) we can obtain a line integral whose value is a vector rather than a scalar, namely,

$$\int_C \mathbf{F}(\mathbf{r})\, dt = \int_a^b \mathbf{F}(\mathbf{r}(t))\, dt = \int_a^b [F_1(\mathbf{r}(t)), \quad F_2(\mathbf{r}(t)), \quad F_3(\mathbf{r}(t))]\, dt. \tag{8}$$

Obviously, a special case of (7) is obtained by taking $F_1 = f$, $F_2 = F_3 = 0$. Then

$$\int_C f(\mathbf{r})\, dt = \int_a^b f(\mathbf{r}(t))\, dt \tag{8*}$$

with C as in (2). The evaluation is similar to that before.

EXAMPLE 5 A Line Integral of the Form (8)

Integrate $\mathbf{F}(\mathbf{r}) = [xy, yz, z]$ along the helix in Example 2.

Solution. $\mathbf{F}(\mathbf{r}(t)) = [\cos t \sin t, 3t \sin t, 3t]$ integrated with respect to t from 0 to 2π gives

$$\int_0^{2\pi} \mathbf{F}(\mathbf{r}(t)\, dt = \left[-\frac{1}{2}\cos^2 t, \quad 3 \sin t - 3t \cos t, \quad \frac{3}{2}t^2\right]\Bigg|_0^{2\pi} = [0, \quad -6\pi, \quad 6\pi^2].$$ ■

Path Dependence

Path dependence of line integrals is practically and theoretically so important that we formulate it as a theorem. And a whole section (Sec. 10.2) will be devoted to conditions under which path dependence does not occur.

THEOREM 2

Path Dependence

The line integral (3) *generally depends not only on* **F** *and on the endpoints A and B of the path, but also on the path itself along which the integral is taken.*

PROOF Almost any example will show this. Take, for instance, the straight segment C_1: $\mathbf{r}_1(t) = [t, t, 0]$ and the parabola C_2: $\mathbf{r}_2(t) = [t, t^2, 0]$ with $0 \leqq t \leqq 1$ (Fig. 223) and integrate $\mathbf{F} = [0, xy, 0]$. Then $\mathbf{F}(\mathbf{r}_1(t)) \cdot \mathbf{r}_1'(t) = t^2$, $\mathbf{F}(\mathbf{r}_2(t)) \cdot \mathbf{r}_2'(t) = 2t^4$, so that integration gives $1/3$ and $2/5$, respectively. ■

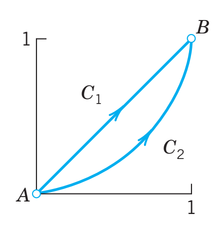


Fig. 223. Proof of Theorem 2

PROBLEM SET 10.1

1. **WRITING PROJECT. From Definite Integrals to Line Integrals.** Write a short report (1–2 pages) with examples on line integrals as generalizations of definite integrals. The latter give the area under a curve. Explain the corresponding geometric interpretation of a line integral.

2–11 LINE INTEGRAL. WORK

Calculate $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ for the given data. If **F** is a force, this gives the work done by the force in the displacement along C. Show the details.

2. $\mathbf{F} = [y^2, -x^2]$, C: $y = 4x^2$ from (0, 0) to (1, 4)
3. **F** as in Prob. 2, C from (0, 0) straight to (1, 4). Compare.
4. $\mathbf{F} = [xy, x^2y^2]$, C from (2, 0) straight to (0, 2)
5. **F** as in Prob. 4, C the quarter-circle from (2, 0) to (0, 2) with center (0, 0)
6. $\mathbf{F} = [x - y, y - z, z - x]$, C: $\mathbf{r} = [2 \cos t, t, 2 \sin t]$ from (2, 0, 0) to $(2, 2\pi, 0)$
7. $\mathbf{F} = [x^2, y^2, z^2]$, C: $\mathbf{r} = [\cos t, \sin t, e^t]$ from (1, 0, 1) to $(1, 0, e^{2\pi})$. Sketch C.
8. $\mathbf{F} = [e^x, \cosh y, \sinh z]$, C: $\mathbf{r} = [t, t^2, t^3]$ from (0, 0, 0) to $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8})$. Sketch C.
9. $\mathbf{F} = [x + y, y + z, z + x]$, C: $\mathbf{r} = [2t, 5t, t]$ from $t = 0$ to 1. Also from $t = -1$ to 1.
10. $\mathbf{F} = [x, -z, 2y]$ from (0, 0, 0) straight to (1, 1, 0), then to (1, 1, 1), back to (0, 0, 0)
11. $\mathbf{F} = [e^{-x}, e^{-y}, e^{-z}]$, C: $\mathbf{r} = [t, t^2, t]$ from (0, 0, 0) to (2, 4, 2). Sketch C.
12. **PROJECT. Change of Parameter. Path Dependence.**

    Consider the integral $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$, where $\mathbf{F} = [xy, -y^2]$.

    (a) One path, several representations. Find the value of the integral when $\mathbf{r} = [\cos t, \sin t]$, $0 \leqq t \leqq \pi/2$. Show that the value remains the same if you set $t = -p$ or $t = p^2$ or apply two other parametric transformations of your own choice.

    (b) Several paths. Evaluate the integral when C: $y = x^n$, thus $\mathbf{r} = [t, t^n]$, $0 \leqq t \leqq 1$, where $n = 1, 2, 3, \cdots$. Note that these infinitely many paths have the same endpoints.

(c) **Limit.** What is the limit in (b) as $n \to \infty$? Can you confirm your result by direct integration without referring to (b)?

(d) Show path dependence with a simple example of your choice involving two paths.

13. ***ML*-Inequality, Estimation of Line Integrals.** Let **F** be a vector function defined on a curve C. Let $|\mathbf{F}|$ be bounded, say, $|\mathbf{F}| \leqq M$ on C, where M is some positive number. Show that

$$\left| \int_C \mathbf{F} \bullet d\mathbf{r} \right| \leqq ML \qquad (L = \text{Length of } C). \tag{9}$$

14. Using (9), find a bound for the absolute value of the work W done by the force $\mathbf{F} = [x^2, y]$ in the displacement from (0, 0) straight to (3, 4). Integrate exactly and compare.

15–20 INTEGRALS (8) AND (8*)

Evaluate them with **F** or f and C as follows.

15. $\mathbf{F} = [y^2, z^2, x^2]$, C: $\mathbf{r} = [3 \cos t, 3 \sin t, 2t]$, $0 \leqq t \leqq 4\pi$

16. $f = 3x + y + 5z$, C: $\mathbf{r} = [t, \cosh t, \sinh t]$, $0 \leqq t \leqq 1$. Sketch C.

17. $\mathbf{F} = [x + y, y + z, z + x]$, C: $\mathbf{r} = [4 \cos t, \sin t, 0]$, $0 \leqq t \leqq \pi$

18. $\mathbf{F} = [y^{1/3}, x^{1/3}, 0]$, C the hypocycloid $\mathbf{r} = [\cos^3 t, \sin^3 t, 0]$, $0 \leqq t \leqq \pi/4$

19. $f = xyz$, C: $\mathbf{r} = [4t, 3t^2, 12t]$, $-2 \leqq t \leqq 2$. Sketch C.

20. $\mathbf{F} = [xz, yz, x^2y^2]$, C: $\mathbf{r} = [t, t, e^t]$, $0 \leqq t \leqq 5$. Sketch C.

10.2 Path Independence of Line Integrals

We want to find out under what conditions, in some domain, a line integral takes on the same value no matter what path of integration is taken (in that domain). As before we consider line integrals

$$\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) \qquad (d\mathbf{r} = [dx, \quad dy, \quad dz]) \tag{1}$$

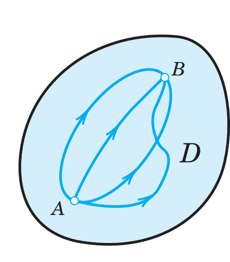


Fig. 224. Path independence

The line integral (1) is said to be **path independent in a domain *D* in space** if for every pair of endpoints A, B in domain D, (1) has the same value for all paths in D that begin at A and end at B. This is illustrated in Fig. 224. (See Sec. 9.6 for "domain.")

Path independence is important. For instance, in mechanics it may mean that we have to do the same amount of work regardless of the path to the mountaintop, be it short and steep or long and gentle. Or it may mean that in releasing an elastic spring we get back the work done in expanding it. Not all forces are of this type—think of swimming in a big round pool in which the water is rotating as in a whirlpool.

We shall follow up with three ideas about path independence. We shall see that path independence of (1) in a domain D holds if and only if:

(*Theorem* 1) $\mathbf{F} = \text{grad} f$, where grad f is the gradient of f as explained in Sec. 9.7.

(*Theorem* 2) Integration around closed curves C in D always gives 0.

(*Theorem* 3) curl $\mathbf{F} = \mathbf{0}$, provided D is simply connected, as defined below.

Do you see that these theorems can help in understanding the examples and counterexample just mentioned?

Let us begin our discussion with the following very practical criterion for path independence.

THEOREM 1

Path Independence

A line integral (1) with continuous F_1, F_2, F_3 in a domain D in space is path independent in D if and only if $\mathbf{F} = [F_1, \quad F_2, \quad F_3]$ is the gradient of some function f in D,

$$\text{(2)} \qquad \mathbf{F} = \operatorname{grad} f, \qquad \text{thus,} \qquad F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}.$$

PROOF **(a)** We assume that (2) holds for some function f in D and show that this implies path independence. Let C be any path in D from any point A to any point B in D, given by $\mathbf{r}(t) = [x(t), \quad y(t), \quad z(t)]$, where $a \leqq t \leqq b$. Then from (2), the chain rule in Sec. 9.6, and (3′) in the last section we obtain

$$\begin{aligned}
\int_C (F_1\,dx + F_2\,dy + F_3\,dz) &= \int_C \left(\frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz\right) \\
&= \int_a^b \left(\frac{\partial f}{\partial x}\frac{dx}{dt} + \frac{\partial f}{\partial y}\frac{dy}{dt} + \frac{\partial f}{\partial z}\frac{dz}{dt}\right)dt \\
&= \int_a^b \frac{df}{dt}\,dt = f[x(t), y(t), z(t)]\Big|_{t=a}^{t=b} \\
&= f(x(b), y(b), z(b)) - f(x(a), y(a), z(a)) \\
&= f(B) - f(A).
\end{aligned}$$

(b) The more complicated proof of the converse, that path independence implies (2) for some f, is given in App. 4. ■

The last formula in part (a) of the proof,

$$\textbf{(3)} \qquad \int_A^B (F_1\,dx + F_2\,dy + F_3\,dz) = f(B) - f(A) \qquad [\mathbf{F} = \operatorname{grad} f]$$

is the analog of the usual formula for definite integrals in calculus,

$$\int_a^b g(x)\,dx = G(x)\Big|_a^b = G(b) - G(a) \qquad [G'(x) = g(x)].$$

Formula (3) should be applied whenever a line integral is independent of path.

Potential theory relates to our present discussion if we remember from Sec. 9.7 that when $\mathbf{F} = \operatorname{grad} f$, then f is called a potential of $\mathbf{F}$. Thus the integral (1) is independent of path in D if and only if $\mathbf{F}$ is the gradient of a potential in D.

EXAMPLE 1 **Path Independence**

Show that the integral $\int_C \mathbf{F} \cdot d\mathbf{r} = \int_C (2x\,dx + 2y\,dy + 4z\,dz)$ is path independent in any domain in space and find its value in the integration from A: (0, 0, 0) to B: (2, 2, 2).

Solution. $\mathbf{F} = [2x, 2y, 4z] = \text{grad} f$, where $f = x^2 + y^2 + 2z^2$ because $\partial f/\partial x = 2x = F_1$, $\partial f/\partial y = 2y = F_2$, $\partial f/\partial z = 4z = F_3$. Hence the integral is independent of path according to Theorem 1, and (3) gives $f(B) - f(A) = f(2, 2, 2) - f(0, 0, 0) = 4 + 4 + 8 = 16$.

If you want to check this, use the most convenient path C: $\mathbf{r}(t) = [t, \quad t, \quad t], 0 \leqq t \leqq 2$, on which $\mathbf{F}(\mathbf{r}(t)) = [2t, \quad 2t, \quad 4t]$, so that $\mathbf{F}(\mathbf{r}(t)) \cdot \mathbf{r}'(t) = 2t + 2t + 4t = 8t$, and integration from 0 to 2 gives $8 \cdot 2^2/2 = 16$.

If you did not see the potential by inspection, use the method in the next example. ■

EXAMPLE 2 **Path Independence. Determination of a Potential**

Evaluate the integral $I = \int_C (3x^2\,dx + 2yz\,dy + y^2\,dz)$ from A: (0, 1, 2) to B: (1, −1, 7) by showing that $\mathbf{F}$ has a potential and applying (3).

Solution. If $\mathbf{F}$ has a potential f, we should have

$$f_x = F_1 = 3x^2, \qquad f_y = F_2 = 2yz, \qquad f_z = F_3 = y^2.$$

We show that we can satisfy these conditions. By integration of f_x and differentiation,

$$\begin{aligned} &f = x^3 + g(y, z), && f_y = g_y = 2yz, && g = y^2 z + h(z), && f = x^3 + y^2 z + h(z) \\ &f_z = y^2 + h' = y^2, && h' = 0 && h = 0, \quad \text{say.} \end{aligned}$$

This gives $f(x, y, z) = x^3 + y^3 z$ and by (3),

$$I = f(1, -1, 7) - f(0, 1, 2) = 1 + 7 - (0 + 2) = 6.$$ ■

Path Independence and Integration Around Closed Curves

The simple idea is that two paths with common endpoints (Fig. 225) make up a single closed curve. This gives almost immediately

THEOREM 2 **Path Independence**

The integral (1) *is path independent in a domain D if and only if its value around every closed path in D is zero.*

PROOF If we have path independence, then integration from A to B along C_1 and along C_2 in Fig. 225 gives the same value. Now C_1 and C_2 together make up a closed curve C, and if we integrate from A along C_1 to B as before, but then in the opposite sense along C_2 back to A (so that this second integral is multiplied by −1), the sum of the two integrals is zero, but this is the integral around the closed curve C.

Conversely, assume that the integral around any closed path C in D is zero. Given any points A and B and any two curves C_1 and C_2 from A to B in D, we see that C_1 with the orientation reversed and C_2 together form a closed path C. By assumption, the integral over C is zero. Hence the integrals over C_1 and C_2, both taken from A to B, must be equal. This proves the theorem. ■

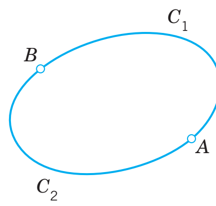


Fig. 225. Proof of Theorem 2

Work. Conservative and Nonconservative (Dissipative) Physical Systems

Recall from the last section that in mechanics, the integral (1) gives the work done by a force $\mathbf{F}$ in the displacement of a body along the curve C. Then Theorem 2 states that work is path independent in D if and only if its value is zero for displacement around every closed path in D. Furthermore, Theorem 1 tells us that this happens if and only if $\mathbf{F}$ is the gradient of a potential in D. In this case, $\mathbf{F}$ and the vector field defined by $\mathbf{F}$ are called **conservative** in D because in this case mechanical energy is conserved; that is, no work is done in the displacement from a point A and back to A. Similarly for the displacement of an electrical charge (an electron, for instance) in a conservative electrostatic field.

Physically, the kinetic energy of a body can be interpreted as the ability of the body to do work by virtue of its motion, and if the body moves in a conservative field of force, after the completion of a round trip the body will return to its initial position with the same kinetic energy it had originally. For instance, the gravitational force is conservative; if we throw a ball vertically up, it will (if we assume air resistance to be negligible) return to our hand with the same kinetic energy it had when it left our hand.

Friction, air resistance, and water resistance always act against the direction of motion. They tend to diminish the total mechanical energy of a system, usually converting it into heat or mechanical energy of the surrounding medium (possibly both). Furthermore, if during the motion of a body, these forces become so large that they can no longer be neglected, then the resultant force $\mathbf{F}$ of the forces acting on the body is no longer conservative. This leads to the following terms. A physical system is called **conservative** if all the forces acting in it are conservative. If this does not hold, then the physical system is called **nonconservative** or **dissipative**.

Path Independence and Exactness of Differential Forms

Theorem 1 relates path independence of the line integral (1) to the gradient and Theorem 2 to integration around closed curves. A third idea (leading to Theorems 3* and 3, below) relates path independence to the exactness of the **differential form** or *Pfaffian form*[1]

$$\mathbf{F} \cdot d\mathbf{r} = F_1\, dx + F_2\, dy + F_3\, dz \tag{4}$$

under the integral sign in (1). This form (4) is called **exact** in a domain D in space if it is the differential

$$df = \frac{\partial f}{\partial x} dx + \frac{\partial f}{\partial y} dy + \frac{\partial f}{\partial z} dz = (\operatorname{grad} f) \cdot d\mathbf{r}$$

of a differentiable function $f(x, y, z)$ everywhere in D, that is, if we have

$$\mathbf{F} \cdot d\mathbf{r} = df.$$

Comparing these two formulas, we see that the form (4) is exact if and only if there is a differentiable function $f(x, y, z)$ in D such that everywhere in D,

$$\mathbf{F} = \operatorname{grad} f, \qquad \text{thus,} \qquad F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}. \tag{5}$$

[1] JOHANN FRIEDRICH PFAFF (1765–1825). German mathematician.

Hence Theorem 1 implies

THEOREM 3*

Path Independence

The integral (1) *is path independent in a domain D in space if and only if the differential form* (4) *has continuous coefficient functions F_1, F_2, F_3 and is exact in D.*

This theorem is of practical importance because it leads to a useful exactness criterion. First we need the following concept, which is of general interest.

A domain D is called **simply connected** if every closed curve in D can be continuously shrunk to any point in D without leaving D.

For example, the interior of a sphere or a cube, the interior of a sphere with finitely many points removed, and the domain between two concentric spheres are simply connected. On the other hand, the interior of a torus, which is a doughnut as shown in Fig. 249 in Sec. 10.6 is not simply connected. Neither is the interior of a cube with one space diagonal removed.

The criterion for exactness (and path independence by Theorem 3*) is now as follows.

THEOREM 3

Criterion for Exactness and Path Independence

Let F_1, F_2, F_3 in the line integral (1),

$$\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz),$$

be continuous and have continuous first partial derivatives in a domain D in space. Then:

(a) *If the differential form* (4) *is exact in D—and thus* (1) *is path independent by Theorem* 3*—, *then in D,*

$$\text{curl } \mathbf{F} = \mathbf{0}; \tag{6}$$

in components (see Sec. 9.9)

$$\frac{\partial F_3}{\partial y} = \frac{\partial F_2}{\partial z}, \qquad \frac{\partial F_1}{\partial z} = \frac{\partial F_3}{\partial x}, \qquad \frac{\partial F_2}{\partial x} = \frac{\partial F_1}{\partial y}. \tag{6'}$$

(b) *If* (6) *holds in D and D is simply connected, then* (4) *is exact in D—and thus* (1) *is path independent by Theorem* 3*.

PROOF **(a)** If (4) is exact in D, then $\mathbf{F} = \text{grad} f$ in D by Theorem 3*, and, furthermore, $\text{curl } \mathbf{F} = \text{curl } (\text{grad} f) = \mathbf{0}$ by (2) in Sec. 9.9, so that (6) holds.

(b) The proof needs "Stokes's theorem" and will be given in Sec. 10.9. ■

Line Integral in the Plane. For $\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy)$ the curl has only one component (the z-component), so that (6′) reduces to the single relation

$$\frac{\partial F_2}{\partial x} = \frac{\partial F_1}{\partial y} \tag{6''}$$

(which also occurs in (5) of Sec. 1.4 on exact ODEs).

EXAMPLE 3 **Exactness and Independence of Path. Determination of a Potential**

Using (6′), show that the differential form under the integral sign of

$$I = \int_C [2xyz^2\,dx + (x^2z^2 + z\cos yz)\,dy + (2x^2yz + y\cos yz)\,dz]$$

is exact, so that we have independence of path in any domain, and find the value of I from A: (0, 0, 1) to B: $(1, \pi/4, 2)$.

Solution. Exactness follows from (6′), which gives

$$(F_3)_y = 2x^2z + \cos yz - yz\sin yz = (F_2)_z$$
$$(F_1)_z = 4xyz = (F_3)_x$$
$$(F_2)_x = 2xz^2 = (F_1)_y.$$

To find f, we integrate F_2 (which is "long," so that we save work) and then differentiate to compare with F_1 and F_3,

$$f = \int F_2\,dy = \int (x^2z^2 + z\cos yz)\,dy = x^2z^2y + \sin yz + g(x, z)$$
$$f_x = 2xz^2y + g_x = F_1 = 2xyz^2, \qquad g_x = 0, \qquad g = h(z)$$
$$f_z = 2x^2zy + y\cos yz + h' = F_3 = 2x^2zy + y\cos yz, \qquad h' = 0.$$

$h' = 0$ implies $h = \text{const}$ and we can take $h = 0$, so that $g = 0$ in the first line. This gives, by (3),

$$f(x, y, z) = x^2yz^2 + \sin yz, \qquad f(B) - f(A) = 1 \cdot \frac{\pi}{4} \cdot 4 + \sin\frac{\pi}{2} - 0 = \pi + 1.$$ ■

The assumption in Theorem 3 that D is simply connected is essential and cannot be omitted. Perhaps the simplest example to see this is the following.

EXAMPLE 4 **On the Assumption of Simple Connectedness in Theorem 3**

Let

(7) $$F_1 = -\frac{y}{x^2 + y^2}, \qquad F_2 = \frac{x}{x^2 + y^2}, \qquad F_3 = 0.$$

Differentiation shows that (6′) is satisfied in any domain of the xy-plane not containing the origin, for example, in the domain D: $\frac{1}{2} < \sqrt{x^2 + y^2} < \frac{3}{2}$ shown in Fig. 226. Indeed, F_1 and F_2 do not depend on z, and $F_3 = 0$, so that the first two relations in (6′) are trivially true, and the third is verified by differentiation:

$$\frac{\partial F_2}{\partial x} = \frac{x^2 + y^2 - x \cdot 2x}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2},$$
$$\frac{\partial F_1}{\partial y} = -\frac{x^2 + y^2 - y \cdot 2y}{(x^2 + y^2)^2} = \frac{y^2 - x^2}{(x^2 + y^2)^2}.$$

Clearly, D in Fig. 226 is not simply connected. If the integral

$$I = \int_C (F_1\,dx + F_2\,dy) = \int_C \frac{-y\,dx + x\,dy}{x^2 + y^2}$$

were independent of path in D, then $I = 0$ on any closed curve in D, for example, on the circle $x^2 + y^2 = 1$. But setting $x = r\cos\theta$, $y = r\sin\theta$ and noting that the circle is represented by $r = 1$, we have

$$x = \cos\theta, \qquad dx = -\sin\theta\,d\theta, \qquad y = \sin\theta, \qquad dy = \cos\theta\,d\theta,$$

so that $-y\,dx + x\,dy = \sin^2\theta\,d\theta + \cos^2\theta\,d\theta = d\theta$ and counterclockwise integration gives

$$I = \int_0^{2\pi} \frac{d\theta}{1} = 2\pi.$$

Since D is not simply connected, we cannot apply Theorem 3 and cannot conclude that I is independent of path in D.

Although $\mathbf{F} = \operatorname{grad} f$, where $f = \arctan(y/x)$ (verify!), we cannot apply Theorem 1 either because the polar angle $f = \theta = \arctan(y/x)$ is not single-valued, as it is required for a function in calculus. ■

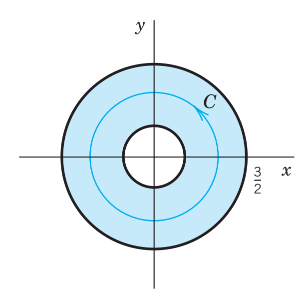


Fig. 226. Example 4

PROBLEM SET 10.2

1. **WRITING PROJECT. Report on Path Independence.** Make a list of the main ideas and facts on path independence and dependence in this section. Then work this list into a report. Explain the definitions and the practical usefulness of the theorems, with illustrative examples of your own. No proofs.

2. **On Example 4.** Does the situation in Example 4 of the text change if you take the domain $0 < \sqrt{x^2 + y^2} < 3/2$?

3–9 PATH INDEPENDENT INTEGRALS

Show that the form under the integral sign is exact in the plane (Probs. 3–4) or in space (Probs. 5–9) and evaluate the integral. Show the details of your work.

3. $\displaystyle\int_{(\pi/2,\,\pi)}^{(\pi,\,0)} \left(\tfrac{1}{2}\cos\tfrac{1}{2}x\cos 2y\,dx - 2\sin\tfrac{1}{2}x\sin 2y\,dy\right)$

4. $\displaystyle\int_{(4,\,0)}^{(6,\,1)} e^{4y}(2x\,dx + 4x^2\,dy)$

5. $\displaystyle\int_{(0,\,0,\,\pi)}^{(2,\,1/2,\,\pi/2)} e^{xy}(y\sin z\,dx + x\sin z\,dy + \cos z\,dz)$

6. $\displaystyle\int_{(0,\,0,\,0)}^{(1,\,1,\,0)} e^{x^2+y^2+z^2}(x\,dx + y\,dy + z\,dz)$

7. $\displaystyle\int_{(0,\,2,\,3)}^{(1,\,1,\,1)} (yz\sinh xz\,dx + \cosh xz\,dy + xy\sinh xz\,dz)$

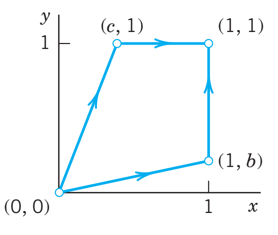
8. $\displaystyle\int_{(5,\,3,\,\pi)}^{(3,\,\pi,\,3)} (\cos yz\,dx - xz\sin yz\,dy - xy\sin yz\,dz)$

9. $\displaystyle\int_{(0,\,1,\,0)}^{(1,\,0,\,1)} (e^x\cosh y\,dx + (e^x\sinh y + e^z\cosh y)\,dy + e^z\sinh y\,dz)$

10. **PROJECT. Path Dependence. (a)** Show that $I = \int_C (x^2y\,dx + 2xy^2\,dy)$ is path dependent in the xy-plane.

(b) Integrate from (0, 0) along the straight-line segment to $(1, b)$, $0 \leqq b \leqq 1$, and then vertically up to (1, 1); see the figure. For which b is I maximum? What is its maximum value?

(c) Integrate I from (0, 0) along the straight-line segment to $(c, 1)$, $0 \leqq c \leqq 1$, and then horizontally to (1, 1). For $c = 1$, do you get the same value as for $b = 1$ in (b)? For which c is I maximum? What is its maximum value?



Project 10. Path Dependence

11. On Example 4. Show that in Example 4 of the text, $\mathbf{F} = \text{grad}(\arctan(y/x))$. Give examples of domains in which the integral is path independent.

12. CAS EXPERIMENT. Extension of Project 10. Integrate $x^2y\,dx + 2xy^2\,dy$ over various circles through the points (0, 0) and (1, 1). Find experimentally the smallest value of the integral and the approximate location of the center of the circle.

13–19 PATH INDEPENDENCE?

Check, and if independent, integrate from (0, 0, 0) to (a, b, c).

13. $2e^{x^2}(x \cos 2y\,dx - \sin 2y\,dy)$

14. $(\sinh xy)(z\,dx - x\,dz)$

15. $x^2y\,dx - 4xy^2\,dy + 8z^2x\,dz$

16. $e^y\,dx + (xe^y - e^z)\,dy - ye^z\,dz$

17. $4y\,dx + z\,dy + (y - 2z)\,dz$

18. $(\cos xy)(yz\,dx + xz\,dy) - 2 \sin xy\,dz$

19. $(\cos(x^2 + 2y^2 + z^2))(2x\,dx + 4y\,dy + 2z\,dz)$

20. Path Dependence. Construct three simple examples in each of which two equations (6′) are satisfied, but the third is not.

10.3 Calculus Review: Double Integrals. *Optional*

This section is optional. Students familiar with double integrals from calculus should skip this review and go on to Sec. 10.4. This section is included in the book to make it reasonably self-contained.

In a definite integral (1), Sec. 10.1, we integrate a function $f(x)$ over an interval (a segment) of the x-axis. In a double integral we integrate a function $f(x, y)$, called the *integrand,* over a closed bounded region[2] R in the xy-plane, whose boundary curve has a unique tangent at almost every point, but may perhaps have finitely many cusps (such as the vertices of a triangle or rectangle).

The definition of the double integral is quite similar to that of the definite integral. We subdivide the region R by drawing parallels to the x- and y-axes (Fig. 227). We number the rectangles that are entirely within R from 1 to n. In each such rectangle we choose a point, say, (x_k, y_k) in the kth rectangle, whose area we denote by ΔA_k. Then we form the sum

$$J_n = \sum_{k=1}^{n} f(x_k, y_k)\,\Delta A_k.$$

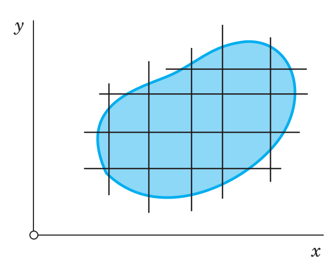


Fig. 227. Subdivision of a region R

[2]A **region** R is a domain (Sec. 9.6) plus, perhaps, some or all of its boundary points. R is **closed** if its **boundary** (all its boundary points) are regarded as belonging to R; and R is **bounded** if it can be enclosed in a circle of sufficiently large radius. A **boundary point** P of R is a point (of R or not) such that every disk with center P contains points of R and also points not of R.

This we do for larger and larger positive integers n in a completely independent manner, but so that the length of the maximum diagonal of the rectangles approaches zero as n approaches infinity. In this fashion we obtain a sequence of real numbers $J_{n_1}, J_{n_2}, \cdots$. Assuming that $f(x, y)$ is continuous in R and R is bounded by finitely many smooth curves (see Sec. 10.1), one can show (see Ref. [GenRef4] in App. 1) that this sequence converges and its limit is independent of the choice of subdivisions and corresponding points (x_k, y_k). This limit is called the **double integral** *of* $f(x, y)$ *over the region* R, and is denoted by

$$\iint_R f(x, y)\, dx\, dy \qquad \text{or} \qquad \iint_R f(x, y)\, dA.$$

Double integrals have properties quite similar to those of definite integrals. Indeed, for any functions f and g of (x, y), defined and continuous in a region R,

$$\iint_R kf\, dx\, dy = k\iint_R f\, dx\, dy \qquad (k \text{ constant})$$

(1)
$$\iint_R (f + g)\, dx\, dy = \iint_R f\, dx\, dy + \iint_R g\, dx\, dy$$

$$\iint_R f\, dx\, dy = \iint_{R_1} f\, dx\, dy + \iint_{R_2} f\, dx\, dy \qquad \text{(Fig. 228)}.$$

Furthermore, if R is simply connected (see Sec. 10.2), then there exists at least one point (x_0, y_0) in R such that we have

(2)
$$\iint_R f(x, y)\, dx\, dy = f(x_0, y_0)A,$$

where A is the area of R. This is called the **mean value theorem** *for double integrals.*

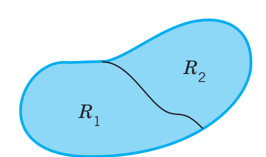


Fig. 228. Formula (1)

Evaluation of Double Integrals by Two Successive Integrations

Double integrals over a region R may be evaluated by two successive integrations. We may integrate first over y and then over x. Then the formula is

(3)
$$\iint_R f(x, y)\, dx\, dy = \int_a^b \left[\int_{g(x)}^{h(x)} f(x, y)\, dy\right] dx \qquad \text{(Fig. 229)}.$$

Here $y = g(x)$ and $y = h(x)$ represent the boundary curve of R (see Fig. 229) and, keeping x constant, we integrate $f(x, y)$ over y from $g(x)$ to $h(x)$. The result is a function of x, and we integrate it from $x = a$ to $x = b$ (Fig. 229).

Similarly, for integrating first over x and then over y the formula is

$$\textbf{(4)} \qquad \iint_R f(x, y)\,dx\,dy = \int_c^d \left[\int_{p(y)}^{q(y)} f(x, y)\,dx \right] dy \qquad \text{(Fig. 230).}$$

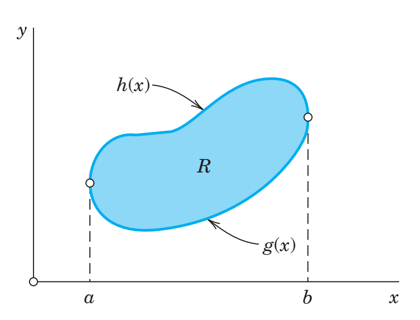


Fig. 229. Evaluation of a double integral

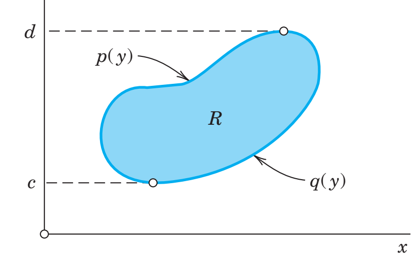


Fig. 230. Evaluation of a double integral

The boundary curve of R is now represented by $x = p(y)$ and $x = q(y)$. Treating y as a constant, we first integrate $f(x, y)$ over x from $p(y)$ to $q(y)$ (see Fig. 230) and then the resulting function of y from $y = c$ to $y = d$.

In (3) we assumed that R can be given by inequalities $a \leqq x \leqq b$ and $g(x) \leqq y \leqq h(x)$. Similarly in (4) by $c \leqq y \leqq d$ and $p(y) \leqq x \leqq q(y)$. If a region R has no such representation, then, in any practical case, it will at least be possible to subdivide R into finitely many portions each of which can be given by those inequalities. Then we integrate $f(x, y)$ over each portion and take the sum of the results. This will give the value of the integral of $f(x, y)$ over the entire region R.

Applications of Double Integrals

Double integrals have various physical and geometric applications. For instance, the **area** A of a region R in the xy-plane is given by the double integral

$$A = \iint_R dx\,dy.$$

The **volume** V beneath the surface $z = f(x, y)$ (>0) and above a region R in the xy-plane is (Fig. 231)

$$V = \iint_R f(x, y)\,dx\,dy$$

because the term $f(x_k, y_k)\Delta A_k$ in J_n at the beginning of this section represents the volume of a rectangular box with base of area ΔA_k and altitude $f(x_k, y_k)$.

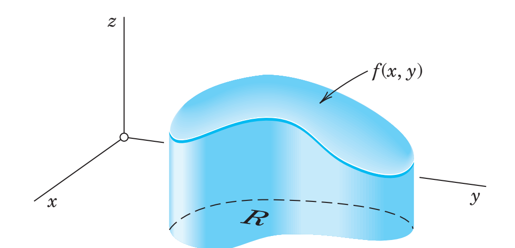


Fig. 231. Double integral as volume

As another application, let $f(x, y)$ be the density ($=$ mass per unit area) of a distribution of mass in the xy-plane. Then the **total mass** M in R is

$$M = \iint_R f(x, y)\, dx\, dy;$$

the **center of gravity** of the mass in R has the coordinates $\bar{x}, \bar{y}$, where

$$\bar{x} = \frac{1}{M} \iint_R x f(x, y)\, dx\, dy \qquad \text{and} \qquad \bar{y} = \frac{1}{M} \iint_R y f(x, y)\, dx\, dy;$$

the **moments of inertia** I_x and I_y of the mass in R about the x- and y-axes, respectively, are

$$I_x = \iint_R y^2 f(x, y)\, dx\, dy, \qquad I_y = \iint_R x^2 f(x, y)\, dx\, dy;$$

and the **polar moment of inertia** I_0 about the origin of the mass in R is

$$I_0 = I_x + I_y = \iint_R (x^2 + y^2) f(x, y)\, dx\, dy.$$

An example is given below.

Change of Variables in Double Integrals. Jacobian

Practical problems often require a change of the variables of integration in double integrals. Recall from calculus that for a definite integral the formula for the change from x to u is

$$\text{(5)} \qquad \int_a^b f(x)\, dx = \int_\alpha^\beta f(x(u)) \frac{dx}{du}\, du.$$

Here we assume that $x = x(u)$ is continuous and has a continuous derivative in some interval $\alpha \leqq u \leqq \beta$ such that $x(\alpha) = a$, $x(\beta) = b$ [or $x(\alpha) = b$, $x(\beta) = a$] and $x(u)$ varies between a and b when u varies between α and β.

The formula for a change of variables in double integrals from x, y to u, v is

$$\textbf{(6)} \qquad \iint_R f(x, y)\, dx\, dy = \iint_{R^*} f(x(u, v), y(u, v)) \left| \frac{\partial(x, y)}{\partial(u, v)} \right| du\, dv;$$

that is, the integrand is expressed in terms of u and v, and $dx\,dy$ is replaced by $du\,dv$ times the absolute value of the **Jacobian**[3]

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}. \tag{7}$$

Here we assume the following. The functions

$$x = x(u, v), \qquad y = y(u, v)$$

effecting the change are continuous and have continuous partial derivatives in some region R^* in the uv-plane such that for every (u, v) in R^* the corresponding point (x, y) lies in R and, conversely, to every (x, y) in R there corresponds one and only one (u, v) in R^*; furthermore, the Jacobian J is either positive throughout R^* or negative throughout R^*. For a proof, see Ref. [GenRef4] in App. 1.

EXAMPLE 1 Change of Variables in a Double Integral

Evaluate the following double integral over the square R in Fig. 232.

$$\iint_R (x^2 + y^2)\,dx\,dy$$

Solution. The shape of R suggests the transformation $x + y = u, x - y = v$. Then $x = \frac{1}{2}(u + v)$, $y = \frac{1}{2}(u - v)$. The Jacobian is

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{vmatrix} = -\frac{1}{2}.$$

R corresponds to the square $0 \leqq u \leqq 2$, $0 \leqq v \leqq 2$. Therefore,

$$\iint_R (x^2 + y^2)\,dx\,dy = \int_0^2 \int_0^2 \frac{1}{2}(u^2 + v^2)\frac{1}{2}\,du\,dv = \frac{8}{3}.$$ ■

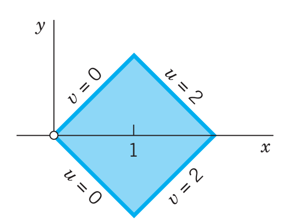


Fig. 232. Region R in Example 1

[3]Named after the German mathematician CARL GUSTAV JACOB JACOBI (1804–1851), known for his contributions to elliptic functions, partial differential equations, and mechanics.

Of particular practical interest are **polar coordinates** r and θ, which can be introduced by setting $x = r\cos\theta,\quad y = r\sin\theta$. Then

$$J = \frac{\partial(x, y)}{\partial(r, \theta)} = \begin{vmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{vmatrix} = r$$

and

$$\text{(8)} \qquad \iint_R f(x, y)\, dx\, dy = \iint_{R^*} f(r\cos\theta, r\sin\theta)\, r\, dr\, d\theta$$

where R^* is the region in the $r\theta$-plane corresponding to R in the xy-plane.

EXAMPLE 2 Double Integrals in Polar Coordinates. Center of Gravity. Moments of Inertia

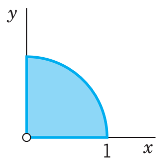


Fig. 233. Example 2

Let $f(x, y) = 1$ be the mass density in the region in Fig. 233. Find the total mass, the center of gravity, and the moments of inertia I_x, I_y, I_0.

Solution. We use the polar coordinates just defined and formula (8). This gives the total mass

$$M = \iint_R dx\, dy = \int_0^{\pi/2}\int_0^1 r\, dr\, d\theta = \int_0^{\pi/2} \frac{1}{2}\, d\theta = \frac{\pi}{4}.$$

The center of gravity has the coordinates

$$\bar{x} = \frac{4}{\pi}\int_0^{\pi/2}\int_0^1 r\cos\theta\, r\, dr\, d\theta = \frac{4}{\pi}\int_0^{\pi/2} \frac{1}{3}\cos\theta\, d\theta = \frac{4}{3\pi} = 0.4244$$

$$\bar{y} = \frac{4}{3\pi} \qquad \text{for reasons of symmetry.}$$

The moments of inertia are

$$I_x = \iint_R y^2\, dx\, dy = \int_0^{\pi/2}\int_0^1 r^2 \sin^2\theta\, r\, dr\, d\theta = \int_0^{\pi/2} \frac{1}{4}\sin^2\theta\, d\theta$$

$$= \int_0^{\pi/2} \frac{1}{8}(1 - \cos 2\theta)\, d\theta = \frac{1}{8}\left(\frac{\pi}{2} - 0\right) = \frac{\pi}{16} = 0.1963$$

$$I_y = \frac{\pi}{16} \quad \text{for reasons of symmetry,} \qquad I_0 = I_x + I_y = \frac{\pi}{8} = 0.3927.$$

Why are $\bar{x}$ and $\bar{y}$ less than $\frac{1}{2}$? ■

This is the end of our review on double integrals. These integrals will be needed in this chapter, beginning in the next section.

PROBLEM SET 10.3

1. Mean value theorem. Illustrate (2) with an example.

2–8 DOUBLE INTEGRALS

Describe the region of integration and evaluate.

2. $\int_0^2 \int_x^{2x} (x + y)^2 \, dy \, dx$

3. $\int_0^3 \int_{-y}^{y} (x^2 + y^2) \, dx \, dy$

4. Prob. 3, order reversed.

5. $\int_0^1 \int_{x^2}^{x} (1 - 2xy) \, dy \, dx$

6. $\int_0^2 \int_0^{y} \sinh(x + y) \, dx \, dy$

7. Prob. 6, order reversed.

8. $\int_0^{\pi/4} \int_0^{\cos y} x^2 \sin y \, dx \, dy$

9–11 VOLUME

Find the volume of the given region in space.

9. The region beneath $z = 4x^2 + 9y^2$ and above the rectangle with vertices (0, 0), (3, 0), (3, 2), (0, 2) in the xy-plane.

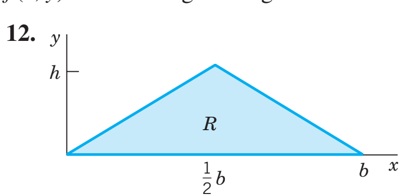
10. The first octant region bounded by the coordinate planes and the surfaces $y = 1 - x^2$, $z = 1 - x^2$. Sketch it.

11. The region above the xy-plane and below the paraboloid $z = 1 - (x^2 + y^2)$.

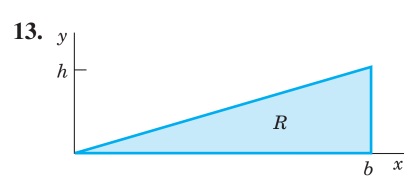
12–16 CENTER OF GRAVITY

Find the center of gravity $(\bar{x}, \bar{y})$ of a mass of density $f(x, y) = 1$ in the given region R.

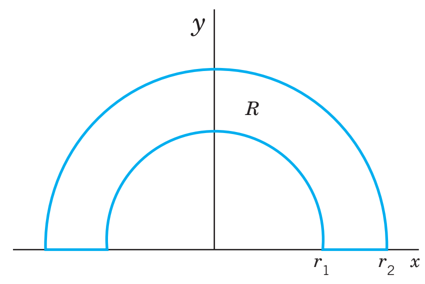
12.



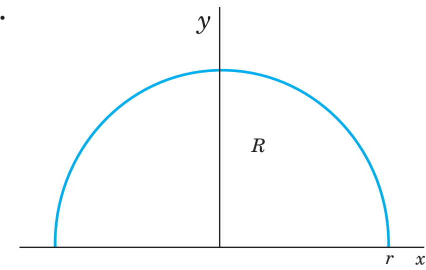
13.



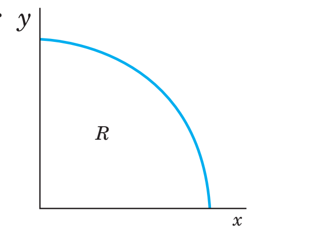
14.



15.



16.



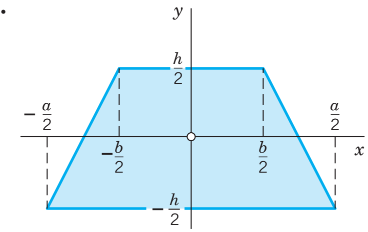
17–20 MOMENTS OF INERTIA

Find I_x, I_y, I_0 of a mass of density $f(x, y) = 1$ in the region R in the figures, which the engineer is likely to need, along with other profiles listed in engineering handbooks.

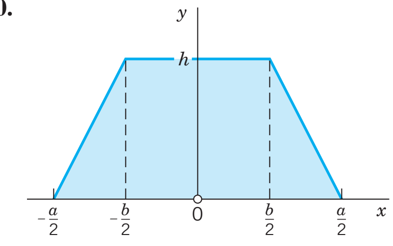
17. R as in Prob. 13.

18. R as in Prob. 12.

19.



20.



10.4 Green's Theorem in the Plane

Double integrals over a plane region may be transformed into line integrals over the boundary of the region and conversely. This is of practical interest because it may simplify the evaluation of an integral. It also helps in theoretical work when we want to switch from one kind of integral to the other. The transformation can be done by the following theorem.

THEOREM 1

Green's Theorem in the Plane[4]
(Transformation between Double Integrals and Line Integrals)

Let R be a closed bounded region (see Sec. 10.3) *in the xy-plane whose boundary C consists of finitely many smooth curves* (see Sec. 10.1). *Let $F_1(x, y)$ and $F_2(x, y)$ be functions that are continuous and have continuous partial derivatives $\partial F_1/\partial y$ and $\partial F_2/\partial x$ everywhere in some domain containing R. Then*

$$\iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dx\,dy = \oint_C (F_1\,dx + F_2\,dy). \tag{1}$$

Here we integrate along the entire boundary C of R in such a sense that R is on the left as we advance in the direction of integration (see Fig. 234).

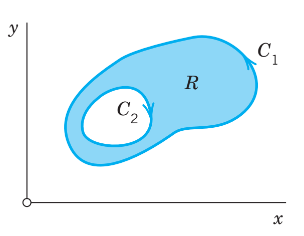


Fig. 234. Region R whose boundary C consists of two parts: C_1 is traversed counterclockwise, while C_2 is traversed clockwise in such a way that R is on the left for both curves

Setting $\mathbf{F} = [F_1, \quad F_2] = F_1\mathbf{i} + F_2\mathbf{j}$ *and using* (1) *in* Sec. 9.9, we obtain (1) in vectorial form,

$$\iint_R (\text{curl}\,\mathbf{F}) \bullet \mathbf{k}\,dx\,dy = \oint_C \mathbf{F} \bullet d\mathbf{r}. \tag{1'}$$

The proof follows after the first example. For $\oint$ see Sec. 10.1.

---

[4]GEORGE GREEN (1793–1841), English mathematician who was self-educated, started out as a baker, and at his death was fellow of Caius College, Cambridge. His work concerned potential theory in connection with electricity and magnetism, vibrations, waves, and elasticity theory. It remained almost unknown, even in England, until after his death.

A "domain containing R" in the theorem guarantees that the assumptions about F_1 and F_2 at boundary points of R are the same as at other points of R.

EXAMPLE 1 Verification of Green's Theorem in the Plane

Green's theorem in the plane will be quite important in our further work. Before proving it, let us get used to it by verifying it for $F_1 = y^2 - 7y$, $F_2 = 2xy + 2x$ and C the circle $x^2 + y^2 = 1$.

Solution. In (1) on the left we get

$$\iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dx\, dy = \iint_R [(2y + 2) - (2y - 7)]\, dx\, dy = 9 \iint_R dx\, dy = 9\pi$$

since the circular disk R has area π.

We now show that the line integral in (1) on the right gives the same value, 9π. We must orient C counterclockwise, say, $\mathbf{r}(t) = [\cos t, \sin t]$. Then $\mathbf{r}'(t) = [-\sin t, \cos t]$, and on C,

$$F_1 = y^2 - 7y = \sin^2 t - 7 \sin t, \qquad F_2 = 2xy + 2x = 2 \cos t \sin t + 2 \cos t.$$

Hence the line integral in (1) becomes, verifying Green's theorem,

$$\oint_C (F_1 x' + F_2 y')\, dt = \int_0^{2\pi} [(\sin^2 t - 7 \sin t)(-\sin t) + 2(\cos t \sin t + \cos t)(\cos t)]\, dt$$

$$= \int_0^{2\pi} (-\sin^3 t + 7 \sin^2 t + 2 \cos^2 t \sin t + 2 \cos^2 t)\, dt$$

$$= 0 + 7\pi - 0 + 2\pi = 9\pi.$$

■

PROOF We prove Green's theorem in the plane, first for a *special region* R that can be represented in both forms

$$a \leqq x \leqq b, \qquad u(x) \leqq y \leqq v(x) \qquad \text{(Fig. 235)}$$

and

$$c \leqq y \leqq d, \qquad p(y) \leqq x \leqq q(y) \qquad \text{(Fig. 236)}$$

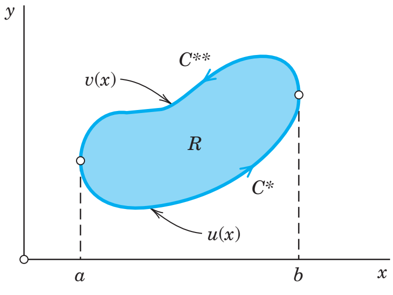


Fig. 235. Example of a special region

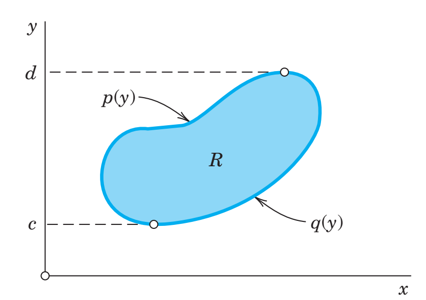


Fig. 236. Example of a special region

Using (3) in the last section, we obtain for the second term on the left side of (1) taken without the minus sign

$$\text{(2)} \qquad \iint_R \frac{\partial F_1}{\partial y}\, dx\, dy = \int_a^b \left[\int_{u(x)}^{v(x)} \frac{\partial F_1}{\partial y}\, dy\right] dx \qquad \text{(see Fig. 235).}$$

(The first term will be considered later.) We integrate the inner integral:

$$\int_{u(x)}^{v(x)} \frac{\partial F_1}{\partial y}\, dy = F_1(x, y)\Big|_{y=u(x)}^{y=v(x)} = F_1[x, v(x)] - F_1[x, u(x)].$$

By inserting this into (2) we find (changing a direction of integration)

$$\begin{aligned}\iint_R \frac{\partial F_1}{\partial y}\, dx\, dy &= \int_a^b F_1[x, v(x)]\, dx - \int_a^b F_1[x, u(x)]\, dx \\ &= -\int_b^a F_1[x, v(x)]\, dx - \int_a^b F_1[x, u(x)]\, dx.\end{aligned}$$

Since $y = v(x)$ represents the curve C^{**} (Fig. 235) and $y = u(x)$ represents C^*, the last two integrals may be written as line integrals over C^{**} and C^* (oriented as in Fig. 235); therefore,

$$\begin{aligned}\iint_R \frac{\partial F_1}{\partial y}\, dx\, dy &= -\int_{C^{**}} F_1(x, y)\, dx - \int_{C^*} F_1(x, y)\, dx \\ &= -\oint_C F_1(x, y)\, dx.\end{aligned} \tag{3}$$

This proves (1) in Green's theorem if $F_2 = 0$.

The result remains valid if C has portions parallel to the y-axis (such as $\tilde{C}$ and $\tilde{\tilde{C}}$ in Fig. 237). Indeed, the integrals over these portions are zero because in (3) on the right we integrate with respect to x. Hence we may add these integrals to the integrals over C^* and C^{**} to obtain the integral over the whole boundary C in (3).

We now treat the first term in (1) on the left in the same way. Instead of (3) in the last section we use (4), and the second representation of the special region (see Fig. 236). Then (again changing a direction of integration)

$$\begin{aligned}\iint_R \frac{\partial F_2}{\partial x}\, dx\, dy &= \int_c^d \left[\int_{p(y)}^{q(y)} \frac{\partial F_2}{\partial x}\, dx\right] dy \\ &= \int_c^d F_2(q(y), y)\, dy + \int_d^c F_2(p(y), y)\, dy \\ &= \oint_C F_2(x, y)\, dy.\end{aligned}$$

Together with (3) this gives (1) and proves Green's theorem for special regions.

We now prove the theorem for a region R that itself is not a special region but can be subdivided into finitely many special regions as shown in Fig. 238. In this case we apply the theorem to each subregion and then add the results; the left-hand members add up to the integral over R while the right-hand members add up to the line integral over C plus

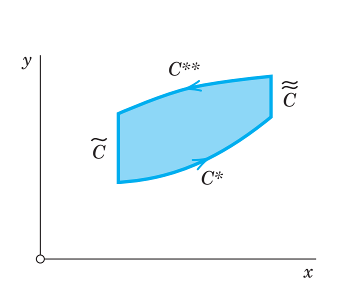


Fig. 237. Proof of Green's theorem

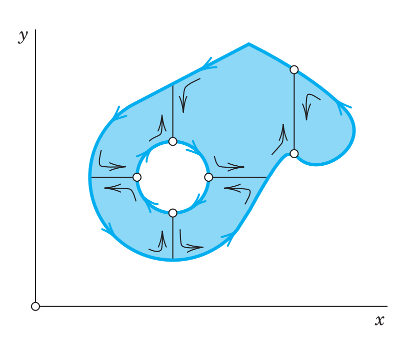


Fig. 238. Proof of Green's theorem

integrals over the curves introduced for subdividing R. The simple ***key observation*** now is that each of the latter integrals occurs twice, taken once in each direction. Hence they cancel each other, leaving us with the line integral over C.

The proof thus far covers all regions that are of interest in practical problems. To prove the theorem for a most general region R satisfying the conditions in the theorem, we must approximate R by a region of the type just considered and then use a limiting process. For details of this see Ref. [GenRef4] in App. 1.

Some Applications of Green's Theorem

EXAMPLE 2 Area of a Plane Region as a Line Integral Over the Boundary

In (1) we first choose $F_1 = 0, F_2 = x$ and then $F_1 = -y, F_2 = 0$. This gives

$$\iint_R dx\,dy = \oint_C x\,dy \qquad \text{and} \qquad \iint_R dx\,dy = -\oint_C y\,dx$$

respectively. The double integral is the area A of R. By addition we have

$$A = \frac{1}{2}\oint_C (x\,dy - y\,dx) \tag{4}$$

where we integrate as indicated in Green's theorem. This interesting formula expresses the area of R in terms of a line integral over the boundary. It is used, for instance, in the theory of certain **planimeters** (mechanical instruments for measuring area). See also Prob. 11.

For an ***ellipse*** $x^2/a^2 + y^2/b^2 = 1$ or $x = a\cos t, y = b\sin t$ we get $x' = -a\sin t, y' = b\cos t$; thus from (4) we obtain the familiar formula for the area of the region bounded by an ellipse,

$$A = \frac{1}{2}\int_0^{2\pi} (xy' - yx')\,dt = \frac{1}{2}\int_0^{2\pi} [ab\cos^2 t - (-ab\sin^2 t)]\,dt = \pi ab. \qquad \blacksquare$$

EXAMPLE 3 Area of a Plane Region in Polar Coordinates

Let r and θ be polar coordinates defined by $x = r\cos\theta, y = r\sin\theta$. Then

$$dx = \cos\theta\,dr - r\sin\theta\,d\theta, \qquad dy = \sin\theta\,dr + r\cos\theta\,d\theta,$$

and (4) becomes a formula that is well known from calculus, namely,

$$A = \frac{1}{2}\oint_C r^2\, d\theta. \tag{5}$$

As an application of (5), we consider the **cardioid** $r = a(1 - \cos\theta)$, where $0 \leqq \theta \leqq 2\pi$ (Fig. 239). We find

$$A = \frac{a^2}{2}\int_0^{2\pi} (1 - \cos\theta)^2\, d\theta = \frac{3\pi}{2}a^2.$$ ■

EXAMPLE 4 Transformation of a Double Integral of the Laplacian of a Function into a Line Integral of Its Normal Derivative

The Laplacian plays an important role in physics and engineering. A first impression of this was obtained in Sec. 9.7, and we shall discuss this further in Chap. 12. At present, let us use Green's theorem for deriving a basic integral formula involving the Laplacian.

We take a function $w(x, y)$ that is continuous and has continuous first and second partial derivatives in a domain of the xy-plane containing a region R of the type indicated in Green's theorem. We set $F_1 = -\partial w/\partial y$ and $F_2 = \partial w/\partial x$. Then $\partial F_1/\partial y$ and $\partial F_2/\partial x$ are continuous in R, and in (1) on the left we obtain

$$\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = \frac{\partial^2 w}{\partial x^2} + \frac{\partial^2 w}{\partial y^2} = \nabla^2 w, \tag{6}$$

the Laplacian of w (see Sec. 9.7). Furthermore, using those expressions for F_1 and F_2, we get in (1) on the right

$$\oint_C (F_1\, dx + F_2\, dy) = \oint_C \left(F_1 \frac{dx}{ds} + F_2 \frac{dy}{ds}\right) ds = \oint_C \left(-\frac{\partial w}{\partial y}\frac{dx}{ds} + \frac{\partial w}{\partial x}\frac{dy}{ds}\right) ds \tag{7}$$

where s is the arc length of C, and C is oriented as shown in Fig. 240. The integrand of the last integral may be written as the dot product

$$(\text{grad } w) \cdot \mathbf{n} = \left[\frac{\partial w}{\partial x}, \frac{\partial w}{\partial y}\right] \cdot \left[\frac{dy}{ds}, -\frac{dx}{ds}\right] = \frac{\partial w}{\partial x}\frac{dy}{ds} - \frac{\partial w}{\partial y}\frac{dx}{ds}. \tag{8}$$

The vector $\mathbf{n}$ is a unit normal vector to C, because the vector $\mathbf{r}'(s) = d\mathbf{r}/ds = [dx/ds, \quad dy/ds]$ is the unit tangent vector of C, and $\mathbf{r}' \cdot \mathbf{n} = 0$, so that $\mathbf{n}$ is perpendicular to $\mathbf{r}'$. Also, $\mathbf{n}$ is directed to the *exterior* of C because in Fig. 240 the positive x-component dx/ds of $\mathbf{r}'$ is the negative y-component of $\mathbf{n}$, and similarly at other points. From this and (4) in Sec. 9.7 we see that the left side of (8) is the derivative of w in the direction of the outward normal of C. This derivative is called the **normal derivative** of w and is denoted by $\partial w/\partial n$; that is, $\partial w/\partial n = (\text{grad } w) \cdot \mathbf{n}$. Because of (6), (7), and (8), Green's theorem gives the desired formula relating the Laplacian to the normal derivative,

$$\iint_R \nabla^2 w\, dx\, dy = \oint_C \frac{\partial w}{\partial n}\, ds. \tag{9}$$

For instance, $w = x^2 - y^2$ satisfies Laplace's equation $\nabla^2 w = 0$. Hence its normal derivative integrated over a closed curve must give 0. Can you verify this directly by integration, say, for the square $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$? ■

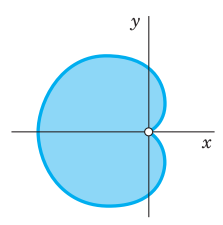


Fig. 239. Cardioid

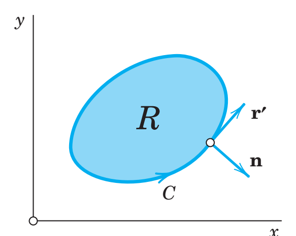


Fig. 240. Example 4

Green's theorem in the plane can be used in both directions, and thus may aid in the evaluation of a given integral by transforming the given integral into another integral that is easier to solve. This is illustrated further in the problem set. Moreover, and perhaps more fundamentally, Green's theorem will be the essential tool in the proof of a very important integral theorem, namely, Stokes's theorem in Sec. 10.9.

PROBLEM SET 10.4

1–10 LINE INTEGRALS: EVALUATION BY GREEN'S THEOREM

Evaluate $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ counterclockwise around the boundary C of the region R by Green's theorem, where

1. $\mathbf{F} = [y, -x]$, C the circle $x^2 + y^2 = 1/4$

2. $\mathbf{F} = [6y^2, 2x - 2y^4]$, R the square with vertices $\pm(2, 2)$, $\pm(2, -2)$

3. $\mathbf{F} = [x^2 e^y, y^2 e^x]$, R the rectangle with vertices (0, 0), (2, 0), (2, 3), (0, 3)

4. $\mathbf{F} = [x \cosh 2y, 2x^2 \sinh 2y]$, $R: x^2 \leqq y \leqq x$

5. $\mathbf{F} = [x^2 + y^2, x^2 - y^2]$, $R: 1 \leqq y \leqq 2 - x^2$

6. $\mathbf{F} = [\cosh y, -\sinh x]$, $R: 1 \leqq x \leqq 3, x \leqq y \leqq 3x$

7. $\mathbf{F} = \text{grad}\,(x^3 \cos^2(xy))$, R as in Prob. 5

8. $\mathbf{F} = [-e^{-x} \cos y, -e^{-x} \sin y]$, R the semidisk $x^2 + y^2 \leqq 16$, $x \geqq 0$

9. $\mathbf{F} = [e^{y/x}, e^y \ln x + 2x]$, $R: 1 + x^4 \leqq y \leqq 2$

10. $\mathbf{F} = [x^2 y^2, -x/y^2]$, $R: 1 \leqq x^2 + y^2 \leqq 4, x \geqq 0,$ $y \geqq x$. Sketch R.

11. CAS EXPERIMENT. Apply (4) to figures of your choice whose area can also be obtained by another method and compare the results.

12. PROJECT. Other Forms of Green's Theorem in the Plane. Let R and C be as in Green's theorem, $\mathbf{r}'$ a unit tangent vector, and $\mathbf{n}$ the outer unit normal vector of C (Fig. 240 in Example 4). Show that (1) may be written

$$\iint_R \text{div}\, \mathbf{F}\, dx\, dy = \oint_C \mathbf{F} \cdot \mathbf{n}\, ds \tag{10}$$

or

$$\iint_R (\text{curl}\ \mathbf{F}) \cdot \mathbf{k}\, dx\, dy = \oint_C \mathbf{F} \cdot \mathbf{r}'\, ds \tag{11}$$

where $\mathbf{k}$ is a unit vector perpendicular to the xy-plane. **Verify** (10) and (11) for $\mathbf{F} = [7x, -3y]$ and C the circle $x^2 + y^2 = 4$ as well as for an example of your own choice.

13–17 INTEGRAL OF THE NORMAL DERIVATIVE

Using (9), find the value of $\int_C \frac{\partial w}{\partial n}\, ds$ taken counterclockwise over the boundary C of the region R.

13. $w = \cosh x$, R the triangle with vertices (0, 0), (4, 2), (0, 2).

14. $w = x^2 y + xy^2$, $R: x^2 + y^2 \leqq 1, x \geqq 0, y \geqq 0$

15. $w = e^x \cos y + xy^3$, $R: 1 \leqq y \leqq 10 - x^2, x \geqq 0$

16. $W = x^2 + y^2$, $C: x^2 + y^2 = 4$. Confirm the answer by direct integration.

17. $w = x^3 - y^3$, $0 \leqq y \leqq x^2$, $|x| \leqq 2$

18. Laplace's equation. Show that for a solution $w(x, y)$ of Laplace's equation $\nabla^2 w = 0$ in a region R with boundary curve C and outer unit normal vector $\mathbf{n}$,

$$\iint_R \left[\left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right] dx\, dy = \oint_C w \frac{\partial w}{\partial n}\, ds. \tag{12}$$

19. Show that $w = e^x \sin y$ satisfies Laplace's equation $\nabla^2 w = 0$ and, using (12), integrate $w(\partial w/\partial n)$ counterclockwise around the boundary curve C of the rectangle $0 \leqq x \leqq 2, 0 \leqq y \leqq 5$.

20. Same task as in Prob. 19 when $w = x^2 + y^2$ and C the boundary curve of the triangle with vertices (0, 0), (1, 0), (0, 1).

10.5 Surfaces for Surface Integrals

Whereas, with line integrals, we integrate over *curves* in space (Secs. 10.1, 10.2), with surface integrals we integrate over *surfaces* in space. Each curve in space is represented by a parametric equation (Secs. 9.5, 10.1). This suggests that we should also find parametric representations for the surfaces in space. This is indeed one of the goals of this section. The surfaces considered are cylinders, spheres, cones, and others. The second goal is to learn about surface normals. Both goals prepare us for Sec. 10.6 on surface integrals. Note that for simplicity, we shall say "surface" also for a portion of a surface.

Representation of Surfaces

Representations of a surface S in xyz-space are

$$z = f(x, y) \qquad \text{or} \qquad g(x, y, z) = 0. \tag{1}$$

For example, $z = +\sqrt{a^2 - x^2 - y^2}$ or $x^2 + y^2 + z^2 - a^2 = 0 \ (z \geqq 0)$ represents a hemisphere of radius a and center 0.

Now for *curves* C in line integrals, it was more practical and gave greater flexibility to use a *parametric* representation $\mathbf{r} = \mathbf{r}(t)$, where $a \leqq t \leqq b$. This is a mapping of the interval $a \leqq t \leqq b$, located on the t-axis, onto the curve C (actually a portion of it) in xyz-space. It maps every t in that interval onto the point of C with position vector $\mathbf{r}(t)$. See Fig. 241A.

Similarly, for surfaces S in surface integrals, it will often be more practical to use a *parametric* representation. Surfaces are *two*-dimensional. Hence we need *two* parameters,

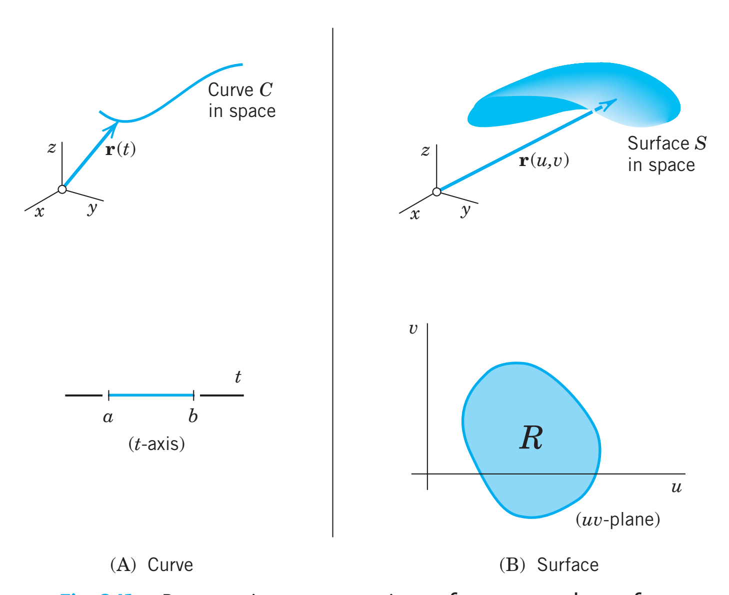


Fig. 241. Parametric representations of a curve and a surface

which we call u and v. Thus a **parametric representation** of a surface S in space is of the form

$$\text{(2)} \qquad \mathbf{r}(u, v) = [x(u, v), y(u, v), z(u, v)] = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

where (u, v) varies in some region R of the uv-plane. This mapping (2) maps every point (u, v) in R onto the point of S with position vector $\mathbf{r}(u, v)$. See Fig. 241B.

EXAMPLE 1 Parametric Representation of a Cylinder

The circular cylinder $x^2 + y^2 = a^2$, $-1 \leqq z \leqq 1$, has radius a, height 2, and the z-axis as axis. A parametric representation is

$$\mathbf{r}(u, v) = [a \cos u, a \sin u, v] = a \cos u\,\mathbf{i} + a \sin u\,\mathbf{j} + v\mathbf{k} \qquad \text{(Fig. 242).}$$

The components of $\mathbf{r}$ are $x = a \cos u$, $y = a \sin u$, $z = v$. The parameters u, v vary in the rectangle R: $0 \leqq u \leqq 2\pi$, $-1 \leqq v \leqq 1$in the uv-plane. The curves $u = \text{const}$ are vertical straight lines. The curves $v = \text{const}$ are parallel circles. The point P in Fig. 242 corresponds to $u = \pi/3 = 60°$, $v = 0.7$. ■

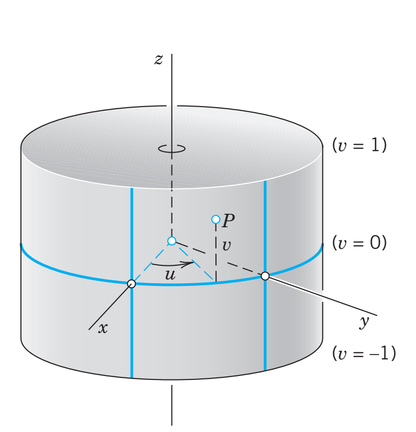


Fig. 242. Parametric representation of a cylinder

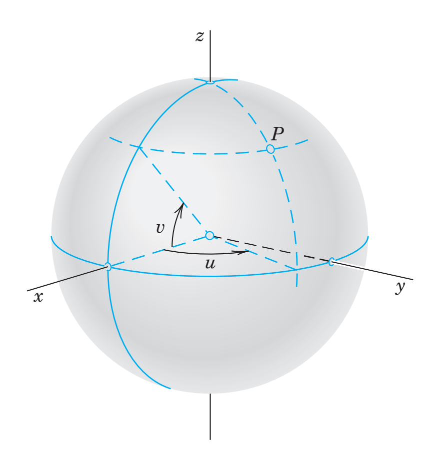


Fig. 243. Parametric representation of a sphere

EXAMPLE 2 Parametric Representation of a Sphere

A sphere $x^2 + y^2 + z^2 = a^2$ can be represented in the form

$$\text{(3)} \qquad \mathbf{r}(u, v) = a \cos v \cos u\,\mathbf{i} + a \cos v \sin u\,\mathbf{j} + a \sin v\,\mathbf{k}$$

where the parameters u, v vary in the rectangle R in the uv-plane given by the inequalities $0 \leqq u \leqq 2\pi$, $-\pi/2 \leqq v \leqq \pi/2$. The components of $\mathbf{r}$ are

$$x = a \cos v \cos u, \qquad y = a \cos v \sin u, \qquad z = a \sin v.$$

The curves $u = \text{const}$ and $v = \text{const}$ are the "meridians" and "parallels" on S (see Fig. 243). *This representation is used in **geography** for measuring the latitude and longitude of points on the globe.*

Another parametric representation of the sphere also used in mathematics is

$$\text{(3*)} \qquad \mathbf{r}(u, v) = a \cos u \sin v\,\mathbf{i} + a \sin u \sin v\,\mathbf{j} + a \cos v\,\mathbf{k}$$

where $0 \leqq u \leqq 2\pi$, $0 \leqq v \leqq \pi$. ■

EXAMPLE 3 Parametric Representation of a Cone

A circular cone $z = \sqrt{x^2 + y^2}, 0 \leqq t \leqq H$ can be represented by

$$\mathbf{r}(u, v) = [u \cos v, u \sin v, u] = u \cos v\,\mathbf{i} + u \sin v\,\mathbf{j} + u\,\mathbf{k},$$

in components $x = u \cos v$, $y = u \sin v$, $z = u$. The parameters vary in the rectangle R: $0 \leqq u \leqq H$, $0 \leqq v \leqq 2\pi$. Check that $x^2 + y^2 = z^2$, as it should be. What are the curves $u =$ const and $v =$ const? ■

Tangent Plane and Surface Normal

Recall from Sec. 9.7 that the tangent vectors of all the curves on a surface S through a point P of S form a plane, called the **tangent plane** of S at P (Fig. 244). Exceptions are points where S has an edge or a cusp (like a cone), so that S cannot have a tangent plane at such a point. Furthermore, a vector perpendicular to the tangent plane is called a **normal vector** of S at P.

Now since S can be given by $\mathbf{r} = \mathbf{r}(u, v)$ in (2), the new idea is that we get a curve C on S by taking a pair of differentiable functions

$$u = u(t), \qquad v = v(t)$$

whose derivatives $u' = du/dt$ and $v' = dv/dt$ are continuous. Then C has the position vector $\tilde{\mathbf{r}}(t) = \mathbf{r}(u(t), v(t))$. By differentiation and the use of the chain rule (Sec. 9.6) we obtain a tangent vector of C on S

$$\tilde{\mathbf{r}}'(t) = \frac{d\tilde{\mathbf{r}}}{dt} = \frac{\partial \mathbf{r}}{\partial u} u' + \frac{\partial \mathbf{r}}{\partial v} v'.$$

Hence the partial derivatives $\mathbf{r}_u$ and $\mathbf{r}_v$ *at P are tangential to S at P.* We assume that they are linearly independent, which geometrically means that the curves $u =$ const and $v =$ const on S intersect at P at a nonzero angle. Then $\mathbf{r}_u$ and $\mathbf{r}_v$ span the tangent plane of S at P. Hence their cross product gives a **normal vector N** of S at P.

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v \neq \mathbf{0}. \tag{4}$$

The corresponding **unit normal vector n** of S at P is (Fig. 244)

$$\mathbf{n} = \frac{1}{|\mathbf{N}|}\mathbf{N} = \frac{1}{|\mathbf{r}_u \times \mathbf{r}_v|}\mathbf{r}_u \times \mathbf{r}_v. \tag{5}$$

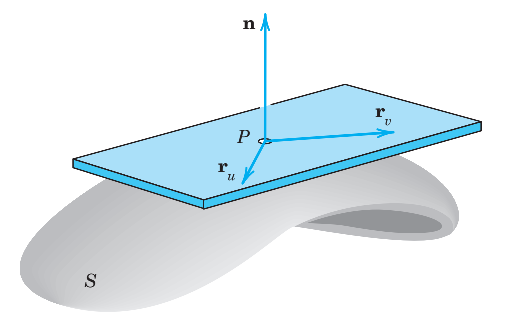


Fig. 244. Tangent plane and normal vector

Also, if S is represented by $g(x, y, z) = 0$, then, by Theorem 2 in Sec. 9.7,

$$\mathbf{n} = \frac{1}{|\text{grad } g|} \text{grad } g. \tag{5*}$$

A surface S is called a **smooth surface** if its surface normal depends continuously on the points of S.

S is called **piecewise smooth** if it consists of finitely many smooth portions.

For instance, a sphere is smooth, and the surface of a cube is piecewise smooth (explain!). We can now summarize our discussion as follows.

THEOREM 1

Tangent Plane and Surface Normal

If a surface S is given by (2) *with continuous $\mathbf{r}_u = \partial\mathbf{r}/\partial u$ and $\mathbf{r}_v = \partial\mathbf{r}/\partial v$ satisfying* (4) *at every point of S, then S has, at every point P, a unique tangent plane passing through P and spanned by $\mathbf{r}_u$ and $\mathbf{r}_v$, and a unique normal whose direction depends continuously on the points of S. A normal vector is given by* (4) *and the corresponding unit normal vector by* (5). (See Fig. 244.)

EXAMPLE 4 **Unit Normal Vector of a Sphere**

From (5*) we find that the sphere $g(x, y, z) = x^2 + y^2 + z^2 - a^2 = 0$ has the unit normal vector

$$\mathbf{n}(x, y, z) = \left[\frac{x}{a}, \frac{y}{a}, \frac{z}{a}\right] = \frac{x}{a}\mathbf{i} + \frac{y}{a}\mathbf{j} + \frac{z}{a}\mathbf{k}.$$

We see that $\mathbf{n}$ has the direction of the position vector $[x, y, z]$ of the corresponding point. Is it obvious that this must be the case? ■

EXAMPLE 5 **Unit Normal Vector of a Cone**

At the apex of the cone $g(x, y, z) = -z + \sqrt{x^2 + y^2} = 0$ in Example 3, the unit normal vector $\mathbf{n}$ becomes undetermined because from (5*) we get

$$\mathbf{n} = \left[\frac{x}{\sqrt{2(x^2 + y^2)}}, \frac{y}{\sqrt{2(x^2 + y^2)}}, \frac{-1}{\sqrt{2}}\right] = \frac{1}{\sqrt{2}}\left(\frac{x}{\sqrt{x^2 + y^2}}\mathbf{i} + \frac{y}{\sqrt{x^2 + y^2}}\mathbf{j} - \mathbf{k}\right).$$ ■

We are now ready to discuss surface integrals and their applications, beginning in the next section.

PROBLEM SET 10.5

1–8 PARAMETRIC SURFACE REPRESENTATION

Familiarize yourself with parametric representations of important surfaces by deriving a representation (1), by finding the **parameter curves** (curves u = const and v = const) of the surface and a normal vector $\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$ of the surface. Show the details of your work.

1. xy-plane $\mathbf{r}(u, v) = (u, \quad v)$ (thus $u\mathbf{i} + v\mathbf{j}$; similarly in Probs. 2–8).
2. xy-plane in polar coordinates $\mathbf{r}(u, v) = [u \cos v, \quad u \sin v]$ (thus $u = r, v = \theta$)
3. Cone $\mathbf{r}(u, v) = [u \cos v, \quad u \sin v, \quad cu]$
4. Elliptic cylinder $\mathbf{r}(u, v) = [a \cos v, \quad b \sin v, \quad u]$
5. Paraboloid of revolution $\mathbf{r}(u, v) = [u \cos v, \quad u \sin v, \quad u^2]$
6. Helicoid $\mathbf{r}(u, v) = [u \cos v, \quad u \sin v, \quad v]$. Explain the name.
7. Ellipsoid $\mathbf{r}(u, v) = [a \cos v \cos u, \quad b \cos v \sin u, \quad c \sin v]$
8. Hyperbolic paraboloid $\mathbf{r}(u, v) = [au \cosh v, \quad bu \sinh v, \quad u^2]$

9. CAS EXPERIMENT. Graphing Surfaces, Dependence on *a*, *b*, *c*. Graph the surfaces in Probs. 3–8. In Prob. 6 generalize the surface by introducing parameters *a, b*. Then find out in Probs. 4 and 6–8 how the shape of the surfaces depends on *a*, *b*, *c*.

10. Orthogonal parameter curves $u = \text{const}$ and $v = \text{const}$ on $\mathbf{r}(u, v)$ occur if and only if $\mathbf{r}_u \bullet \mathbf{r}_v = 0$. Give examples. Prove it.

11. Satisfying (4). Represent the paraboloid in Prob. 5 so that $\tilde{\mathbf{N}}(0, 0) \neq \mathbf{0}$ and show $\tilde{\mathbf{N}}$.

12. Condition (4). Find the points in Probs. 1–8 at which (4) $\mathbf{N} \neq \mathbf{0}$ does not hold. Indicate whether this results from the shape of the surface or from the choice of the representation.

13. Representation $z = f(x, y)$. Show that $z = f(x, y)$ or $g = z - f(x, y) = 0$ can be written ($f_u = \partial f/\partial u$, etc.)

(6) $$\mathbf{r}(u, v) = [u, \quad v, \quad f(u, v)] \quad \text{and}$$
$$\mathbf{N} = \text{grad } g = [-f_u, \quad -f_v, \quad 1].$$

14–19 DERIVE A PARAMETRIC REPRESENTATION

Find a normal vector. The answer gives *one* representation; there are many. Sketch the surface and parameter curves.

14. Plane $4x + 3y + 2z = 12$

15. Cylinder of revolution $(x - 2)^2 + (y + 1)^2 = 25$

16. Ellipsoid $x^2 + y^2 + \frac{1}{9}z^2 = 1$

17. Sphere $x^2 + (y + 2.8)^2 + (z - 3.2)^2 = 2.25$

18. Elliptic cone $z = \sqrt{x^2 + 4y^2}$

19. Hyperbolic cylinder $x^2 - y^2 = 1$

20. PROJECT. Tangent Planes $T(P)$ will be less important in our work, but you should know how to represent them.

(a) If S: $\mathbf{r}(u, v)$, then $T(P)$: $(\mathbf{r}^* - \mathbf{r} \quad \mathbf{r}_u \quad \mathbf{r}_v) = 0$ (a scalar triple product) or
$$\mathbf{r}^*(p, q) = \mathbf{r}(P) + p\mathbf{r}_u(P) + q\mathbf{r}_v(P).$$

(b) If S: $g(x, y, z) = 0$, then
$$T(P): (\mathbf{r}^* - \mathbf{r}(P)) \cdot \nabla g = 0.$$

(c) If S: $z = f(x, y)$, then
$$T(P): z^* - z = (x^* - x)f_x(P) + (y^* - y)f_y(P).$$

Interpret (a)–(c) geometrically. Give two examples for (a), two for (b), and two for (c).

10.6 Surface Integrals

To define a surface integral, we take a surface S, given by a parametric representation as just discussed,

$$\text{(1)} \qquad \mathbf{r}(u, v) = [x(u, v), y(u, v), z(u, v)] = x(u, v)\mathbf{i} + y(u, v)\mathbf{j} + z(u, v)\mathbf{k}$$

where (u, v) varies over a region R in the uv-plane. We assume S to be piecewise smooth (Sec. 10.5), so that S has a normal vector

$$\text{(2)} \qquad \mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v \qquad \text{and unit normal vector} \qquad \mathbf{n} = \frac{1}{|\mathbf{N}|}\mathbf{N}$$

at every point (except perhaps for some edges or cusps, as for a cube or cone). For a given vector function $\mathbf{F}$ we can now define the **surface integral** over S by

$$\textbf{(3)} \qquad \iint_S \mathbf{F} \bullet \mathbf{n} \, dA = \iint_R \mathbf{F}(\mathbf{r}(u, v)) \bullet \mathbf{N}(u, v) \, du \, dv.$$

Here $\mathbf{N} = |\mathbf{N}|\mathbf{n}$ by (2), and $|\mathbf{N}| = |\mathbf{r}_u \times \mathbf{r}_v|$ is the area of the parallelogram with sides $\mathbf{r}_u$ and $\mathbf{r}_v$, by the definition of cross product. Hence

$$\textbf{(3*)} \qquad \mathbf{n} \, dA = \mathbf{n}|\mathbf{N}| \, du \, dv = \mathbf{N} \, du \, dv.$$

And we see that $dA = |\mathbf{N}| du \, dv$ is the element of area of S.

Also $\mathbf{F} \cdot \mathbf{n}$ is the normal component of $\mathbf{F}$. This integral arises naturally in flow problems, where it gives the **flux** across S when $\mathbf{F} = \rho \mathbf{v}$. Recall, from Sec. 9.8, that the flux across S is the mass of fluid crossing S per unit time. Furthermore, ρ is the density of the fluid and $\mathbf{v}$ the velocity vector of the flow, as illustrated by Example 1 below. We may thus call the surface integral (3) the **flux integral**.

We can write (3) in components, using $\mathbf{F} = [F_1, \ F_2, \ F_3]$, $\mathbf{N} = [N_1, \ N_2, \ N_3]$, and $\mathbf{n} = [\cos\alpha, \cos\beta, \cos\gamma]$. Here, α, β, γ are the angles between $\mathbf{n}$ and the coordinate axes; indeed, for the angle between $\mathbf{n}$ and $\mathbf{i}$, formula (4) in Sec. 9.2 gives $\cos\alpha = \mathbf{n} \cdot \mathbf{i}/|\mathbf{n}||\mathbf{i}| = \mathbf{n} \cdot \mathbf{i}$, and so on. We thus obtain from (3)

$$\textbf{(4)} \qquad \begin{aligned} \iint_S \mathbf{F} \cdot \mathbf{n} \, dA &= \iint_S (F_1 \cos\alpha + F_2 \cos\beta + F_3 \cos\gamma) \, dA \\ &= \iint_R (F_1 N_1 + F_2 N_2 + F_3 N_3) \, du \, dv. \end{aligned}$$

In (4) we can write $\cos\alpha \, dA = dy \, dz$, $\cos\beta \, dA = dz \, dx$, $\cos\gamma \, dA = dx \, dy$. Then (4) becomes the following integral for the flux:

$$\textbf{(5)} \qquad \iint_S \mathbf{F} \cdot \mathbf{n} \, dA = \iint_S (F_1 \, dy \, dz + F_2 \, dz \, dx + F_3 \, dx \, dy).$$

We can use this formula to evaluate surface integrals by converting them to double integrals over regions in the coordinate planes of the xyz-coordinate system. But we must carefully take into account the orientation of S (the choice of $\mathbf{n}$). We explain this for the integrals of the F_3-terms,

$$(5') \qquad \iint_S F_3 \cos\gamma \, dA = \iint_S F_3 \, dx \, dy.$$

If the surface S is given by $z = h(x, y)$ with (x, y) varying in a region $\bar{R}$ in the xy-plane, and if S is oriented so that $\cos\gamma > 0$, then $(5')$ gives

$$(5'') \qquad \iint_S F_3 \cos\gamma \, dA = +\iint_{\bar{R}} F_3(x, y, h(x, y)) \, dx \, dy.$$

But if $\cos\gamma < 0$, the integral on the right of $(5'')$ gets a minus sign in front. This follows if we note that the element of area $dx \, dy$ in the xy-plane is the projection $|\cos\gamma| \, dA$ of the element of area dA of S; and we have $\cos\gamma = +|\cos\gamma|$ when $\cos\gamma > 0$, but $\cos\gamma = -|\cos\gamma|$ when $\cos\gamma < 0$. Similarly for the other two terms in (5). At the same time, this justifies the notations in (5).

Other forms of surface integrals will be discussed later in this section.

EXAMPLE 1 Flux Through a Surface

Compute the flux of water through the parabolic cylinder S: $y = x^2$, $0 \leqq x \leqq 2$, $0 \leqq z \leqq 3$ (Fig. 245) if the velocity vector is $\mathbf{v} = \mathbf{F} = [3z^2, \ 6, \ 6xz]$, speed being measured in meters/sec. (Generally, $\mathbf{F} = \rho\mathbf{v}$, but water has the density $\rho = 1 \text{ g/cm}^3 = 1 \text{ ton/m}^3$.)

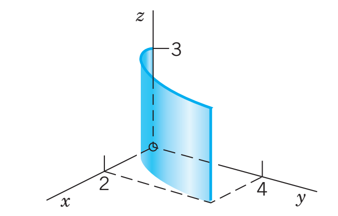


Fig. 245. Surface S in Example 1

Solution. Writing $x = u$ and $z = v$, we have $y = x^2 = u^2$. Hence a representation of S is

$$S: \quad \mathbf{r} = [u, u^2, v] \qquad (0 \leqq u \leqq 2, 0 \leqq v \leqq 3).$$

By differentiation and by the definition of the cross product,

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [1, \quad 2u, \quad 0] \times [0, \quad 0, \quad 1] = [2u, \quad -1, \quad 0].$$

On S, writing simply $\mathbf{F}(S)$ for $\mathbf{F}[\mathbf{r}(u, v)]$, we have $\mathbf{F}(S) = [3v^2, 6, 6uv]$. Hence $\mathbf{F}(S) \bullet \mathbf{N} = 6uv^2 - 6$. By integration we thus get from (3) the flux

$$\iint_S \mathbf{F} \bullet \mathbf{n}\, dA = \int_0^3 \int_0^2 (6uv^2 - 6)\, du\, dv = \int_0^3 (3u^2v^2 - 6u)\Big|_{u=0}^{2} dv$$

$$= \int_0^3 (12v^2 - 12)\, dv = (4v^3 - 12v)\Big|_{v=0}^{3} = 108 - 36 = 72 \ [\text{m}^3/\text{sec}]$$

or 72,000 liters/sec. Note that the y-component of $\mathbf{F}$ is positive (equal to 6), so that in Fig. 245 the flow goes from left to right.

Let us confirm this result by (5). Since

$$\mathbf{N} = |\mathbf{N}|\mathbf{n} = |\mathbf{N}|[\cos\alpha, \quad \cos\beta, \quad \cos\gamma] = [2u, \quad -1, \quad 0] = [2x, \quad -1, \quad 0]$$

we see that $\cos\alpha > 0$, $\cos\beta < 0$, and $\cos\gamma = 0$. Hence the second term of (5) on the right gets a minus sign, and the last term is absent. This gives, in agreement with the previous result,

$$\iint_S \mathbf{F} \bullet \mathbf{n}\, dA = \int_0^3 \int_0^4 3z^2\, dy\, dz - \int_0^2 \int_0^3 6\, dz\, dx = \int_0^3 4(3z^2)\, dz - \int_0^2 6 \cdot 3\, dx = 4 \cdot 3^3 - 6 \cdot 3 \cdot 2 = 72. \quad ■$$

EXAMPLE 2 **Surface Integral**

Evaluate (3) when $\mathbf{F} = [x^2, 0, 3y^2]$ and S is the portion of the plane $x + y + z = 1$ in the first octant (Fig. 246).

Solution. Writing $x = u$ and $y = v$, we have $z = 1 - x - y = 1 - u - v$. Hence we can represent the plane $x + y + z = 1$ in the form $\mathbf{r}(u, v) = [u, v, 1 - u - v]$. We obtain the first-octant portion S of this plane by restricting $x = u$ and $y = v$ to the projection R of S in the xy-plane. R is the triangle bounded by the two coordinate axes and the straight line $x + y = 1$, obtained from $x + y + z = 1$ by setting $z = 0$. Thus $0 \leqq x \leqq 1 - y$, $0 \leqq y \leqq 1$.

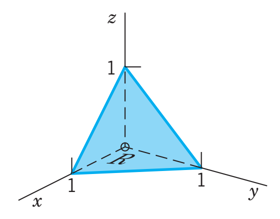


Fig. 246. Portion of a plane in Example 2

By inspection or by differentiation,

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [1, 0, -1] \times [0, 1, -1] = [1, 1, 1].$$

Hence $\mathbf{F}(S) \cdot \mathbf{N} = [u^2, 0, 3v^2] \cdot [1, 1, 1] = u^2 + 3v^2$. By (3),

$$\iint_S \mathbf{F} \cdot \mathbf{n}\, dA = \iint_R (u^2 + 3v^2)\, du\, dv = \int_0^1 \int_0^{1-v} (u^2 + 3v^2)\, du\, dv$$

$$= \int_0^1 \left[\frac{1}{3}(1 - v)^3 + 3v^2(1 - v)\right] dv = \frac{1}{3}.$$

Orientation of Surfaces

From (3) or (4) we see that the value of the integral depends on the choice of the unit normal vector $\mathbf{n}$. (Instead of $\mathbf{n}$ we could choose $-\mathbf{n}$.) We express this by saying that such an integral is an *integral over an* **oriented surface** S, that is, over a surface S on which we have chosen one of the two possible unit normal vectors in a continuous fashion. (For a piecewise smooth surface, this needs some further discussion, which we give below.) If we change the orientation of S, this means that we replace $\mathbf{n}$ with $-\mathbf{n}$. Then each component of $\mathbf{n}$ in (4) is multiplied by -1, so that we have

THEOREM 1

Change of Orientation in a Surface Integral

The replacement of $\mathbf{n}$ *by* $-\mathbf{n}$ (*hence of* $\mathbf{N}$ *by* $-\mathbf{N}$) *corresponds to the multiplication of the integral in* (3) *or* (4) *by* -1.

In practice, how do we make such a change of $\mathbf{N}$ happen, if S is given in the form (1)? The easiest way is to interchange u and v, because then $\mathbf{r}_u$ becomes $\mathbf{r}_v$ and conversely, so that $\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v$ becomes $\mathbf{r}_v \times \mathbf{r}_u = -\mathbf{r}_u \times \mathbf{r}_v = -\mathbf{N}$, as wanted. Let us illustrate this.

EXAMPLE 3 **Change of Orientation in a Surface Integral**

In Example 1 we now represent S by $\tilde{\mathbf{r}} = [v, v^2, u]$, $0 \leqq v \leqq 2$, $0 \leqq u \leqq 3$. Then

$$\tilde{\mathbf{N}} = \tilde{\mathbf{r}}_u \times \tilde{\mathbf{r}}_v = [0, 0, 1] \times [1, 2v, 0] = [-2v, 1, 0].$$

For $\mathbf{F} = [3z^2, 6, 6xz]$ we now get $\tilde{\mathbf{F}}(S) = [3u^2, 6, 6uv]$. Hence $\tilde{\mathbf{F}}(S) \cdot \tilde{\mathbf{N}} = -6u^2v + 6$ and integration gives the old result times -1,

$$\iint_R \tilde{\mathbf{F}}(S) \cdot \tilde{\mathbf{N}}\, dv\, du = \int_0^3 \int_0^2 (-6u^2v + 6)\, dv\, du = \int_0^3 (-12u^2 + 12)\, du = -72.$$

Orientation of Smooth Surfaces

A smooth surface S (see Sec. 10.5) is called **orientable** if the positive normal direction, when given at an arbitrary point P_0 of S, can be continued in a unique and continuous way to the entire surface. In many practical applications, the surfaces are smooth and thus orientable.

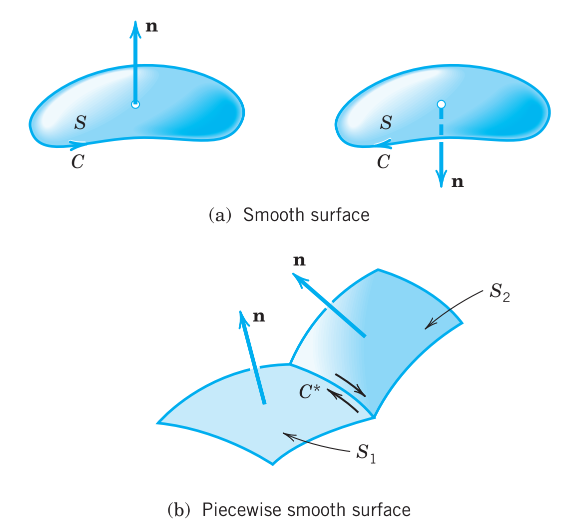


Fig. 247. Orientation of a surface

Orientation of Piecewise Smooth Surfaces

Here the following idea will do it. For a *smooth* orientable surface S with boundary curve C we may associate with each of the two possible orientations of S an orientation of C, as shown in Fig. 247a. Then a *piecewise smooth* surface is called **orientable** if we can orient each smooth piece of S so that along each curve C^* which is a common boundary of two pieces S_1 and S_2 the positive direction of C^* relative to S_1 is opposite to the direction of C^* relative to S_2. See Fig. 247b for two adjacent pieces; note the arrows along C^*.

Theory: Nonorientable Surfaces

A sufficiently small piece of a smooth surface is always orientable. This may not hold for entire surfaces. A well-known example is the **Möbius strip**,[5] shown in Fig. 248. To make a model, take the rectangular paper in Fig. 248, make a half-twist, and join the short sides together so that A goes onto A, and B onto B. At P_0 take a normal vector pointing, say, to the *left*. Displace it along C to the right (in the lower part of the figure) around the strip until you return to P_0 and see that you get a normal vector pointing to the *right*, opposite to the given one. See also Prob. 17.

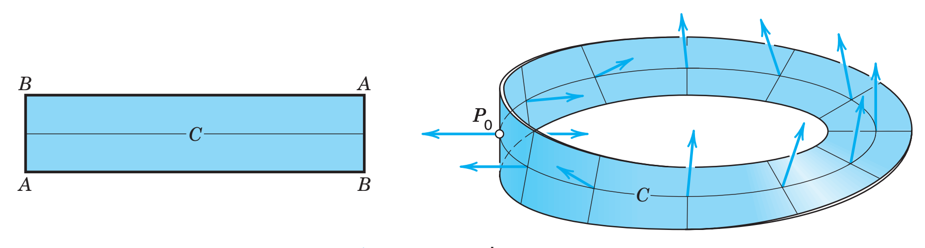


Fig. 248. Möbius strip

[5]AUGUST FERDINAND MÖBIUS (1790–1868), German mathematician, student of Gauss, known for his work in surface theory, geometry, and complex analysis (see Sec. 17.2).

Surface Integrals Without Regard to Orientation

Another type of surface integral is

$$\textbf{(6)} \qquad \iint_S G(\mathbf{r})\, dA = \iint_R G(\mathbf{r}(u, v))|\mathbf{N}(u, v)|\, du\, dv.$$

Here $dA = |\mathbf{N}|\, du\, dv = |\mathbf{r}_u \times \mathbf{r}_v|\, du\, dv$ is the element of area of the surface S represented by (1) and we disregard the orientation.

We shall need later (in Sec. 10.9) the **mean value theorem** for surface integrals, which states that if R in (6) is simply connected (see Sec. 10.2) and $G(\mathbf{r})$ is continuous in a domain containing R, then there is a point (u_0, v_0) in R such that

$$(7) \qquad \iint_S G(\mathbf{r})\, dA = G(\mathbf{r}(u_0, v_0))A \qquad (A = \text{Area of } S).$$

As for applications, if $G(\mathbf{r})$ is the mass density of S, then (6) is the total mass of S. If $G = 1$, then (6) gives the **area** $A(S)$ of S,

$$\textbf{(8)} \qquad A(S) = \iint_S dA = \iint_R |\mathbf{r}_u \times \mathbf{r}_v|\, du\, dv.$$

Examples 4 and 5 show how to apply (8) to a sphere and a torus. The final example, Example 6, explains how to calculate moments of inertia for a surface.

EXAMPLE 4 Area of a Sphere

For a sphere $\mathbf{r}(u, v) = [a \cos v \cos u, \quad a \cos v \sin u, \quad a \sin v], \quad 0 \leqq u \leqq 2\pi, \quad -\pi/2 \leqq v \leqq \pi/2$ [see (3) in Sec. 10.5], we obtain by direct calculation (verify!)

$$\mathbf{r}_u \times \mathbf{r}_v = [a^2 \cos^2 v \cos u, \qquad a^2 \cos^2 v \sin u, \qquad a^2 \cos v \sin v].$$

Using $\cos^2 u + \sin^2 u = 1$ and then $\cos^2 v + \sin^2 v = 1$, we obtain

$$|\mathbf{r}_u \times \mathbf{r}_v| = a^2(\cos^4 v \cos^2 u + \cos^4 v \sin^2 u + \cos^2 v \sin^2 v)^{1/2} = a^2|\cos v|.$$

With this, (8) gives the familiar formula (note that $|\cos v| = \cos v$ when $-\pi/2 \leqq v \leqq \pi/2$)

$$A(S) = a^2 \int_{-\pi/2}^{\pi/2} \int_0^{2\pi} |\cos v|\, du\, dv = 2\pi a^2 \int_{-\pi/2}^{\pi/2} \cos v\, dv = 4\pi a^2. \qquad \blacksquare$$

EXAMPLE 5 Torus Surface (Doughnut Surface): Representation and Area

A *torus surface* S is obtained by rotating a circle C about a straight line L in space so that C does not intersect or touch L but its plane always passes through L. If L is the z-axis and C has radius b and its center has distance a ($> b$) from L, as in Fig. 249, then S can be represented by

$$\mathbf{r}(u, v) = (a + b \cos v) \cos u\, \mathbf{i} + (a + b \cos v) \sin u\, \mathbf{j} + b \sin v\, \mathbf{k}$$

where $0 \leqq u \leqq 2\pi, 0 \leqq v \leqq 2\pi$. Thus

$$\mathbf{r}_u = -(a + b \cos v) \sin u\, \mathbf{i} + (a + b \cos v) \cos u\, \mathbf{j}$$

$$\mathbf{r}_v = -b \sin v \cos u\, \mathbf{i} - b \sin v \sin u\, \mathbf{j} + b \cos v\, \mathbf{k}$$

$$\mathbf{r}_u \times \mathbf{r}_v = b(a + b \cos v)(\cos u \cos v\, \mathbf{i} + \sin u \cos v\, \mathbf{j} + \sin v\, \mathbf{k}).$$

Hence $|\mathbf{r}_u \times \mathbf{r}_v| = b(a + b\cos v)$, and (8) gives the total area of the torus,

$$A(S) = \int_0^{2\pi}\int_0^{2\pi} b(a + b\cos v)\, du\, dv = 4\pi^2 ab. \tag{9}$$

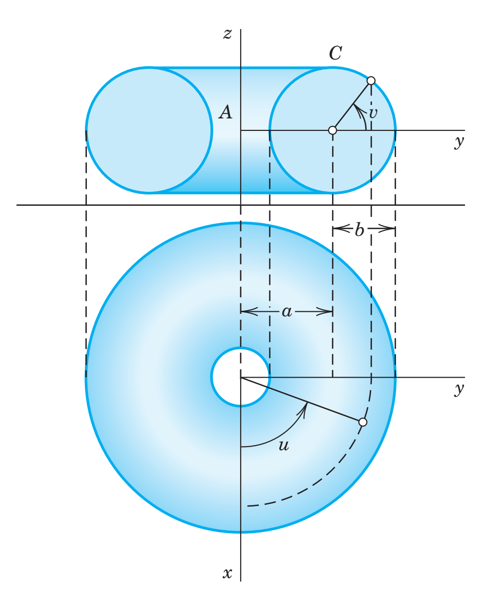


Fig. 249. Torus in Example 5

EXAMPLE 6 Moment of Inertia of a Surface

Find the moment of inertia I of a spherical lamina $S{:}\ = x^2 + y^2 + z^2 = a^2$ of constant mass density and total mass M about the z-axis.

Solution. If a mass is distributed over a surface S and $\mu(x, y, z)$ is the density of the mass (= mass per unit area), then the moment of inertia I of the mass with respect to a given axis L is defined by the surface integral

$$I = \iint_S \mu D^2\, dA \tag{10}$$

where $D(x, y, z)$ is the distance of the point (x, y, z) from L. Since, in the present example, μ is constant and S has the area $A = 4\pi a^2$, we have $\mu = M/A = M/(4\pi a^2)$.

For S we use the same representation as in Example 4. Then $D^2 = x^2 + y^2 = a^2\cos^2 v$. Also, as in that example, $dA = a^2 \cos v\, du\, dv$. This gives the following result. [In the integration, use $\cos^3 v = \cos v\,(1 - \sin^2 v)$.]

$$I = \iint_S \mu D^2\, dA = \frac{M}{4\pi a^2}\int_{-\pi/2}^{\pi/2}\int_0^{2\pi} a^4 \cos^3 v\, du\, dv = \frac{Ma^2}{2}\int_{-\pi/2}^{\pi/2} \cos^3 v\, dv = \frac{2Ma^2}{3}.$$

Representations $z = f(x, y)$. If a surface S is given by $z = f(x, y)$, then setting $u = x$, $v = y$, $\mathbf{r} = [u, v, f]$ gives

$$|\mathbf{N}| = |\mathbf{r}_u \times \mathbf{r}_v| = |[1, 0, f_u] \times [0, 1, f_v]| = |[-f_u, -f_v, 1]| = \sqrt{1 + f_u^2 + f_v^2}$$

and, since $f_u = f_x, f_v = f_y$, formula (6) becomes

$$\iint_S G(\mathbf{r})\, dA = \iint_{R^*} G(x, y, f(x, y)) \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dx\, dy. \tag{11}$$

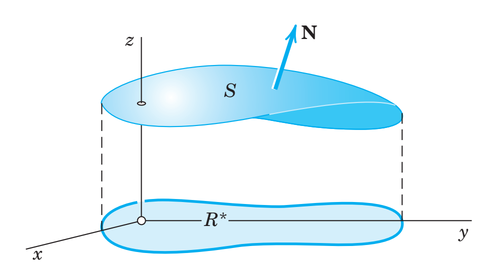


Fig. 250. Formula (11)

Here R^* is the projection of S into the xy-plane (Fig. 250) and the normal vector $\mathbf{N}$ on S points *up*. If it points *down*, the integral on the right is preceded by a minus sign.

From (11) with $G = 1$ we obtain for the **area** $A(S)$ of S: $z = f(x, y)$ the formula

$$A(S) = \iint_{R^*} \sqrt{1 + \left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2}\, dx\, dy \qquad (12)$$

where R^* is the projection of S into the xy-plane, as before.

PROBLEM SET 10.6

1–10 FLUX INTEGRALS (3) $\int_S \mathbf{F} \cdot \mathbf{n}\, dA$

Evaluate the integral for the given data. Describe the kind of surface. Show the details of your work.

1. $\mathbf{F} = [-x^2, y^2, 0]$, S: $\mathbf{r} = [u, v, 3u - 2v]$, $0 \leqq u \leqq 1.5$, $-2 \leqq v \leqq 2$
2. $\mathbf{F} = [e^y, e^x, 1]$, S: $x + y + z = 1$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
3. $\mathbf{F} = [0, x, 0]$, S: $x^2 + y^2 + z^2 = 1$, $x \geqq 0$, $y \geqq 0$, $z \geqq 0$
4. $\mathbf{F} = [e^y, -e^z, e^x]$, S: $x^2 + y^2 = 25$, $x \geqq 0$, $y \geqq 0$, $0 \leqq z \leqq 2$
5. $\mathbf{F} = [x, y, z]$, S: $\mathbf{r} = [u \cos v, u \sin v, u^2]$, $0 \leqq u \leqq 4$, $-\pi \leqq v \leqq \pi$
6. $\mathbf{F} = [\cosh y, 0, \sinh x]$, S: $z = x + y^2$, $0 \leqq y \leqq x$, $0 \leqq x \leqq 1$
7. $\mathbf{F} = [0, \sin y, \cos z]$, S the cylinder $x = y^2$, where $0 \leqq y \leqq \pi/4$ and $0 \leqq z \leqq y$
8. $\mathbf{F} = [\tan xy, x, y]$, S: $y^2 + z^2 = 1$, $2 \leqq x \leqq 5$, $y \geqq 0$, $z \geqq 0$
9. $\mathbf{F} = [0, \sinh z, \cosh x]$, S: $x^2 + z^2 = 4$, $0 \leqq x \leqq 1/\sqrt{2}$, $0 \leqq y \leqq 5$, $z \geqq 0$
10. $\mathbf{F} = [y^2, x^2, z^4]$, S: $z = 4\sqrt{x^2 + y^2}$, $0 \leqq z \leqq 8$, $y \geqq 0$
11. **CAS EXPERIMENT. Flux Integral.** Write a program for evaluating surface integrals (3) that prints intermediate results ($\mathbf{F}$, $\mathbf{F} \cdot \mathbf{N}$, the integral over one of the two variables). Can you obtain experimentally some rules on functions and surfaces giving integrals that can be evaluated by the usual methods of calculus? Make a list of positive and negative results.

12–16 SURFACE INTEGRALS (6) $\iint_S G(\mathbf{r})\, dA$

Evaluate these integrals for the following data. Indicate the kind of surface. Show the details.

12. $G = \cos x + \sin x$, S the portion of $x + y + z = 1$ in the first octant
13. $G = x + y + z$, $z = x + 2y$, $0 \leqq x \leqq \pi$, $0 \leqq y \leqq x$
14. $G = ax + by + cz$, S: $x^2 + y^2 + z^2 = 1$, $y = 0$, $z = 0$
15. $G = (1 + 9xz)^{3/2}$, S: $\mathbf{r} = [u, v, u^3]$, $0 \leqq u \leqq 1$, $-2 \leqq v \leqq 2$
16. $G = \arctan(y/x)$, S: $z = x^2 + y^2$, $1 \leqq z \leqq 9$, $x \geqq 0$, $y \geqq 0$
17. **Fun with Möbius.** Make Möbius strips from long slim rectangles R of grid paper (graph paper) by pasting the short sides together after giving the paper a half-twist. In each case count the number of parts obtained by cutting along lines parallel to the edge. **(a)** Make R three squares wide and cut until you reach the beginning. **(b)** Make R four squares wide. Begin cutting one square away from the edge until you reach the beginning. Then cut the portion that is still two squares wide. **(c)** Make

R five squares wide and cut similarly. **(d)** Make R six squares wide and cut. Formulate a conjecture about the number of parts obtained.

18. Gauss "Double Ring" (See Möbius, *Works* **2**, 518–559). Make a paper cross (Fig. 251) into a "double ring" by joining opposite arms along their outer edges (without twist), one ring below the plane of the cross and the other above. Show experimentally that one can choose any four boundary points A, B, C, D and join A and C as well as B and D by two nonintersecting curves. What happens if you cut along the two curves? If you make a half-twist in each of the two rings and then cut? (Cf. E. Kreyszig, Proc. CSHPM 13 (2000), 23–43.)

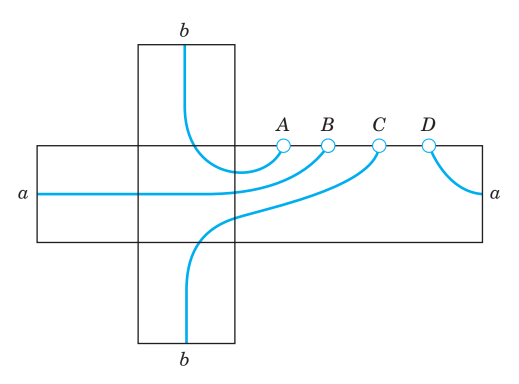


Fig. 251. **Problem 18.** Gauss "Double Ring"

APPLICATIONS

19. Center of gravity. Justify the following formulas for the mass M and the center of gravity $(\bar{x}, \bar{y}, \bar{z})$ of a lamina S of density (mass per unit area) $\sigma(x, y, z)$ in space:

$$M = \iint_S \sigma\, dA, \quad \bar{x} = \frac{1}{M}\iint_S x\sigma\, dA,$$

$$\bar{y} = \frac{1}{M}\iint_S y\sigma\, dA, \quad \bar{z} = \frac{1}{M}\iint_S z\sigma\, dA.$$

20. Moments of inertia. Justify the following formulas for the moments of inertia of the lamina in Prob. 19 about the x-, y-, and z-axes, respectively:

$$I_x = \iint_S (y^2 + z^2)\sigma\, dA, \quad I_y = \iint_S (x^2 + z^2)\sigma\, dA,$$

$$I_z = \iint_S (x^2 + y^2)\sigma\, dA.$$

21. Find a formula for the moment of inertia of the lamina in Prob. 20 about the line $y = x, z = 0$.

22–23 Find the moment of inertia of a lamina S of density 1 about an axis B, where

22. $S: x^2 + y^2 = 1,\ 0 \leqq z \leqq h$, B: the line $z = h/2$ in the xz-plane

23. $S: x^2 + y^2 = z^2,\ 0 \leqq z \leqq h$, B: the z-axis

24. Steiner's theorem.[6] If I_B is the moment of inertia of a mass distribution of total mass M with respect to a line B through the center of gravity, show that its moment of inertia I_K with respect to a line K, which is parallel to B and has the distance k from it is

$$I_K = I_B + k^2 M.$$

25. Using Steiner's theorem, find the moment of inertia of a mass of density 1 on the sphere $S: x^2 + y^2 + z^2 = 1$ about the line $K: x = 1, y = 0$ from the moment of inertia of the mass about a suitable line B, which you must first calculate.

26. TEAM PROJECT. First Fundamental Form of S. Given a surface $S: \mathbf{r}(u, v)$, the differential form

$$ds^2 = E\, du^2 + 2F\, du\, dv + G\, dv^2 \tag{13}$$

with coefficients (in standard notation, unrelated to F, G elsewhere in this chapter)

$$E = \mathbf{r}_u \cdot \mathbf{r}_u, \quad F = \mathbf{r}_u \cdot \mathbf{r}_v, \quad G = \mathbf{r}_v \cdot \mathbf{r}_v \tag{14}$$

is called the **first fundamental form** of S. This form is basic because it permits us to calculate lengths, angles, and areas on S. To show this prove (a)–(c):

(a) For a curve $C: u = u(t), v = v(t), a \leqq t \leqq b$, on S, formulas (10), Sec. 9.5, and (14) give the length

$$\begin{aligned} l &= \int_a^b \sqrt{\mathbf{r}'(t) \cdot \mathbf{r}'(t)}\, dt \\ &= \int_a^b \sqrt{Eu'^2 + 2Fu'v' + Gv'^2}\, dt. \end{aligned} \tag{15}$$

(b) The angle γ between two intersecting curves $C_1: u = g(t), v = h(t)$ and $C_2: u = p(t), v = q(t)$ on $S: \mathbf{r}(u, v)$ is obtained from

$$\cos\gamma = \frac{\mathbf{a} \cdot \mathbf{b}}{|\mathbf{a}||\mathbf{b}|} \tag{16}$$

where $\mathbf{a} = \mathbf{r}_u g' + \mathbf{r}_v h'$ and $\mathbf{b} = \mathbf{r}_u p' + \mathbf{r}_v q'$ are tangent vectors of C_1 and C_2.

[6]JACOB STEINER (1796–1863), Swiss geometer, born in a small village, learned to write only at age 14, became a pupil of Pestalozzi at 18, later studied at Heidelberg and Berlin and, finally, because of his outstanding research, was appointed professor at Berlin University.

(c) The square of the length of the normal vector **N** can be written

$$|\mathbf{N}|^2 = |\mathbf{r}_u \times \mathbf{r}_v|^2 = EG - F^2, \tag{17}$$

so that formula (8) for the area $A(S)$ of S becomes

$$\begin{aligned} A(S) &= \iint_S dA = \iint_R |\mathbf{N}|\, du\, dv \\ &= \iint_R \sqrt{EG - F^2}\, du\, dv. \end{aligned} \tag{18}$$

(d) For polar coordinates $u\ (= r)$ and $v\ (= \theta)$ defined by $x = u \cos v$, $y = u \sin v$ we have $E = 1$, $F = 0$, $G = u^2$, so that

$$ds^2 = du^2 + u^2\, dv^2 = dr^2 + r^2\, d\theta^2.$$

Calculate from this and (18) the area of a disk of radius a.

(e) Find the first fundamental form of the torus in Example 5. Use it to calculate the area A of the torus. Show that A can also be obtained by the **theorem of Pappus**,[7] which states that the area of a surface of revolution equals the product of the length of a meridian C and the length of the path of the center of gravity of C when C is rotated through the angle 2π.

(f) Calculate the first fundamental form for the usual representations of important surfaces of your own choice (cylinder, cone, etc.) and apply them to the calculation of lengths and areas on these surfaces.

10.7 Triple Integrals. Divergence Theorem of Gauss

In this section we discuss another "big" integral theorem, the divergence theorem, which transforms surface integrals into triple integrals. So let us begin with a review of the latter.

A **triple integral** is an integral of a function $f(x, y, z)$ taken over a closed bounded, three-dimensional region T in space. (Note that "closed" and "bounded" are defined in the same way as in footnote 2 of Sec. 10.3, with "sphere" substituted for "circle"). We subdivide T by planes parallel to the coordinate planes. Then we consider those boxes of the subdivision that lie entirely inside T, and number them from 1 to n. Here each box consists of a rectangular parallelepiped. In each such box we choose an arbitrary point, say, (x_k, y_k, z_k) in box k. The volume of box k we denote by ΔV_k. We now form the sum

$$J_n = \sum_{k=1}^{n} f(x_k, y_k, z_k)\, \Delta V_k.$$

This we do for larger and larger positive integers n arbitrarily but so that the maximum length of all the edges of those n boxes approaches zero as n approaches infinity. This gives a sequence of real numbers $J_{n_1}, J_{n_2}, \cdots$. We assume that $f(x, y, z)$ is continuous in a domain containing T, and T is bounded by finitely many *smooth surfaces* (see Sec. 10.5). Then it can be shown (see Ref. [GenRef4] in App. 1) that the sequence converges to a limit that is independent of the choice of subdivisions and corresponding points

[7]PAPPUS OF ALEXANDRIA (about A.D. 300), Greek mathematician. The theorem is also called Guldin's theorem. HABAKUK GULDIN (1577–1643) was born in St. Gallen, Switzerland, and later became professor in Graz and Vienna.

(x_k, y_k, z_k). This limit is called the **triple integral** *of* $f(x, y, z)$ *over the region* T and is denoted by

$$\iiint_T f(x, y, z)\, dx\, dy\, dz \qquad \text{or by} \qquad \iiint_T f(x, y, z)\, dV.$$

Triple integrals can be evaluated by three successive integrations. This is similar to the evaluation of double integrals by two successive integrations, as discussed in Sec. 10.3. Example 1 below explains this.

Divergence Theorem of Gauss

Triple integrals can be transformed into surface integrals over the boundary surface of a region in space and conversely. Such a transformation is of practical interest because one of the two kinds of integral is often simpler than the other. It also helps in establishing fundamental equations in fluid flow, heat conduction, etc., as we shall see. The transformation is done by the *divergence theorem,* which involves the **divergence** of a vector function $\mathbf{F} = [F_1, F_2, F_3] = F_1\mathbf{i} + F_2\mathbf{j} + F_3\mathbf{k}$, namely,

(1) $$\operatorname{div} \mathbf{F} = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z} \qquad \text{(Sec. 9.8).}$$

THEOREM 1

Divergence Theorem of Gauss
(Transformation Between Triple and Surface Integrals)

Let T *be a closed bounded region in space whose boundary is a piecewise smooth orientable surface* S. *Let* $\mathbf{F}(x, y, z)$ *be a vector function that is continuous and has continuous first partial derivatives in some domain containing* T. *Then*

(2) $$\iiint_T \operatorname{div} \mathbf{F}\, dV = \iint_S \mathbf{F} \cdot \mathbf{n}\, dA.$$

In components *of* $\mathbf{F} = [F_1, \quad F_2, \quad F_3]$ *and of the outer unit normal vector* $\mathbf{n} = [\cos\alpha, \quad \cos\beta, \quad \cos\gamma]$ *of* S (as in Fig. 253), *formula* (2) *becomes*

(2*) $$\begin{aligned}\iiint_T \left(\frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}\right) dx\, dy\, dz &= \iint_S (F_1 \cos\alpha + F_2 \cos\beta + F_3 \cos\gamma)\, dA \\ &= \iint_S (F_1\, dy\, dz + F_2\, dz\, dx + F_3\, dx\, dy).\end{aligned}$$

"Closed bounded region" is explained above, "piecewise smooth orientable" in Sec. 10.5, and "domain containing T" in footnote 4, Sec. 10.4, for the two-dimensional case.

Before we prove the theorem, let us show a standard application.

EXAMPLE 1 Evaluation of a Surface Integral by the Divergence Theorem

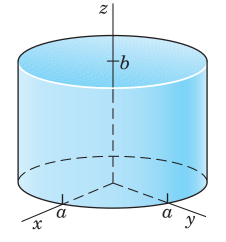


Fig. 252. Surface S in Example 1

Before we prove the theorem, let us show a typical application. Evaluate

$$I = \iint_S (x^3\, dy\, dz + x^2 y\, dz\, dx + x^2 z\, dx\, dy)$$

where S is the closed surface in Fig. 252 consisting of the cylinder $x^2 + y^2 = a^2$ $(0 \leqq z \leqq b)$ and the circular disks $z = 0$ and $z = b$ $(x^2 + y^2 \leqq a^2)$.

Solution. $F_1 = x^3$, $F_2 = x^2 y$, $F_3 = x^2 z$. Hence $\operatorname{div} \mathbf{F} = 3x^2 + x^2 + x^2 = 5x^2$. The form of the surface suggests that we introduce polar coordinates r, θ defined by $x = r\cos\theta$, $y = r\sin\theta$ (thus cylindrical coordinates r, θ, z). Then the volume element is $dx\, dy\, dz = r\, dr\, d\theta\, dz$, and we obtain

$$I = \iiint_T 5x^2\, dx\, dy\, dz = \int_{z=0}^{b} \int_{\theta=0}^{2\pi} \int_{r=0}^{a} (5r^2 \cos^2\theta)\, r\, dr\, d\theta\, dz$$

$$= 5 \int_{z=0}^{b} \int_{\theta=0}^{2\pi} \frac{a^4}{4} \cos^2\theta\, d\theta\, dz = 5 \int_{z=0}^{b} \frac{a^4 \pi}{4}\, dz = \frac{5\pi}{4} a^4 b.$$ ■

PROOF We prove the divergence theorem, beginning with the first equation in (2*). This equation is true if and only if the integrals of each component on both sides are equal; that is,

$$\iiint_T \frac{\partial F_1}{\partial x}\, dx\, dy\, dz = \iint_S F_1 \cos\alpha\, dA, \tag{3}$$

$$\iiint_T \frac{\partial F_2}{\partial y}\, dx\, dy\, dz = \iint_S F_2 \cos\beta\, dA, \tag{4}$$

$$\iiint_T \frac{\partial F_3}{\partial z}\, dx\, dy\, dz = \iint_S F_3 \cos\gamma\, dA. \tag{5}$$

We first prove (5) for a ***special region*** T that is bounded by a piecewise smooth orientable surface S and has the property that any straight line parallel to any one of the coordinate axes and intersecting T has at most *one* segment (or a single point) in common with T. This implies that T can be represented in the form

$$g(x, y) \leqq z \leqq h(x, y) \tag{6}$$

where (x, y) varies in the orthogonal projection $\bar{R}$ of T in the xy-plane. Clearly, $z = g(x, y)$ represents the "bottom" S_2 of S (Fig. 253), whereas $z = h(x, y)$ represents the "top" S_1 of S, and there may be a remaining vertical portion S_3 of S. (The portion S_3 may degenerate into a curve, as for a sphere.)

To prove (5), we use (6). Since **F** is continuously differentiable in some domain containing T, we have

$$\iiint_T \frac{\partial F_3}{\partial z}\,dx\,dy\,dz = \iint_{\bar{R}} \left[\int_{g(x,y)}^{h(x,y)} \frac{\partial F_3}{\partial z}\,dz\right] dx\,dy. \tag{7}$$

Integration of the inner integral $[\cdots]$ gives $F_3[x, y, h(x, y)] - F_3[x, y, g(x, y)]$. Hence the triple integral in (7) equals

$$\iint_{\bar{R}} F_3[x, y, h(x, y)]\,dx\,dy - \iint_{\bar{R}} F_3[x, y, g(x, y)]\,dx\,dy. \tag{8}$$

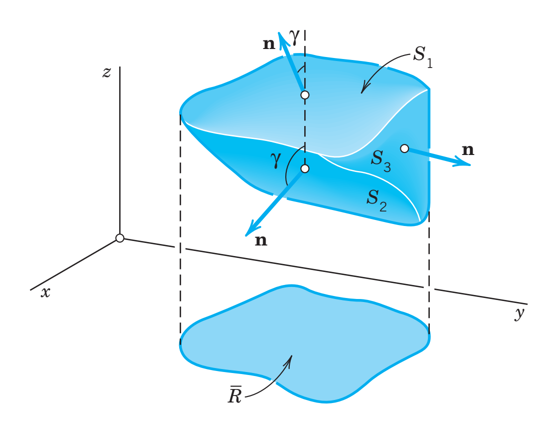


Fig. 253. Example of a special region

But the same result is also obtained by evaluating the right side of (5); that is [see also the last line of (2*)],

$$\iint_S F_3 \cos\gamma\,dA = \iint_S F_3\,dx\,dy$$

$$= +\iint_{\bar{R}} F_3[x, y, h(x, y)]\,dx\,dy - \iint_{\bar{R}} F_3[x, y, g(x, y)]\,dx\,dy,$$

where the first integral over $\bar{R}$ gets a plus sign because $\cos\gamma > 0$ on S_1 in Fig. 253 [as in (5″), Sec. 10.6], and the second integral gets a minus sign because $\cos\gamma < 0$ on S_2. This proves (5).

The relations (3) and (4) now follow by merely relabeling the variables and using the fact that, by assumption, T has representations similar to (6), namely,

$$\tilde{g}(y, z) \leqq x \leqq \tilde{h}(y, z) \qquad \text{and} \qquad \tilde{\tilde{g}}(z, x) \leqq y \leqq \tilde{\tilde{h}}(z, x).$$

This proves the first equation in (2*) for special regions. It implies (2) because the left side of (2*) is just the definition of the divergence, and the right sides of (2) and of the first equation in (2*) are equal, as was shown in the first line of (4) in the last section. Finally, equality of the right sides of (2) and (2*), last line, is seen from (5) in the last section.

This establishes the divergence theorem for special regions.

For any region T that can be subdivided into *finitely many* special regions by means of auxiliary surfaces, the theorem follows by adding the result for each part separately. This procedure is analogous to that in the proof of Green's theorem in Sec. 10.4. The surface integrals over the auxiliary surfaces cancel in pairs, and the sum of the remaining surface integrals is the surface integral over the whole boundary surface S of T; the triple integrals over the parts of T add up to the triple integral over T.

The divergence theorem is now proved for any bounded region that is of interest in practical problems. The extension to a most general region T of the type indicated in the theorem would require a certain limit process; this is similar to the situation in the case of Green's theorem in Sec. 10.4. ■

EXAMPLE 2 Verification of the Divergence Theorem

Evaluate $\iint_S (7x\mathbf{i} - z\mathbf{k}) \bullet \mathbf{n}\, dA$ over the sphere $S: x^2 + y^2 + z^2 = 4$ (a) by (2), (b) directly.

Solution. **(a)** $\operatorname{div} \mathbf{F} = \operatorname{div} [7x, 0, -z] = \operatorname{div} [7x\mathbf{i} - z\mathbf{k}] = 7 - 1 = 6$. *Answer:* $6 \cdot (\frac{4}{3})\pi \cdot 2^3 = 64\pi$.

(b) We can represent S by (3), Sec. 10.5 (with $a = 2$), and we shall use $\mathbf{n}\, dA = \mathbf{N}\, du\, dv$ [see (3*), Sec. 10.6]. Accordingly,

$$S: \quad \mathbf{r} = [2 \cos v \cos u, \quad 2 \cos v \sin u, \quad 2 \sin u] \qquad .$$

Then

$$\mathbf{r}_u = [-2 \cos v \sin u, \quad 2 \cos v \cos u, \quad 0]$$

$$\mathbf{r}_v = [-2 \sin v \cos u, \quad -2 \sin v \sin u, \quad 2 \cos v]$$

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = [4 \cos^2 v \cos u, \quad 4 \cos^2 v \sin u, \quad 4 \cos v \sin v].$$

Now on S we have $x = 2 \cos v \cos u$, $z = 2 \sin v$, so that $\mathbf{F} = [7x, 0, -z]$ becomes on S

$$\mathbf{F}(S) = [14 \cos v \cos u, \quad 0, \quad -2 \sin v]$$

and

$$\begin{aligned} \mathbf{F}(S) \bullet \mathbf{N} &= (14 \cos v \cos u) \cdot 4 \cos^2 v \cos u + (-2 \sin v) \cdot 4 \cos v \sin v \\ &= 56 \cos^3 v \cos^2 u - 8 \cos v \sin^2 v. \end{aligned}$$

On S we have to integrate over u from 0 to 2π. This gives

$$\pi \cdot 56 \cos^3 v - 2\pi \cdot 8 \cos v \sin^2 v.$$

The integral of $\cos v \sin^2 v$ equals $(\sin^3 v)/3$, and that of $\cos^3 v = \cos v\,(1 - \sin^2 v)$ equals $\sin v - (\sin^3 v)/3$. On S we have $-\pi/2 \leqq v \leqq \pi/2$, so that by substituting these limits we get

$$56\pi(2 - \tfrac{2}{3}) - 16\pi \cdot \tfrac{2}{3} = 64\pi$$

as hoped for. To see the point of Gauss's theorem, compare the amounts of work. ■

Coordinate Invariance of the Divergence. The divergence (1) is defined in terms of coordinates, but we can use the divergence theorem to show that div $\mathbf{F}$ has a meaning independent of coordinates.

For this purpose we first note that triple integrals have properties quite similar to those of double integrals in Sec. 10.3. In particular, the **mean value theorem for triple integrals** asserts that for any continuous function $f(x, y, z)$ in a bounded and simply connected region T there is a point $Q: (x_0, y_0, z_0)$ in T such that

$$\iiint_T f(x, y, z)\, dV = f(x_0, y_0, z_0)\, V(T) \qquad (V(T) = \text{volume of } T). \tag{9}$$

In this formula we interchange the two sides, divide by $V(T)$, and set $f = \text{div}\,\mathbf{F}$. Then by the divergence theorem we obtain for the divergence an integral over the boundary surface $S(T)$ of T,

$$\text{(10)} \qquad \text{div}\,\mathbf{F}(x_0, y_0, z_0) = \frac{1}{V(T)} \iiint_T \text{div}\,\mathbf{F}\, dV = \frac{1}{V(T)} \iint_{S(T)} \mathbf{F} \cdot \mathbf{n}\, dA.$$

We now choose a point $P: (x_1, y_1, z_1)$ in T and let T shrink down onto P so that the maximum distance $d(T)$ of the points of T from P goes to zero. Then $Q: (x_0, y_0, z_0)$ must approach P. Hence (10) becomes

$$\textbf{(11)} \qquad \text{div}\,\mathbf{F}(P) = \lim_{d(T)\to 0} \frac{1}{V(T)} \iint_{S(T)} \mathbf{F} \cdot \mathbf{n}\, dA.$$

This proves

THEOREM 2

Invariance of the Divergence

The divergence of a vector function **F** *with continuous first partial derivatives in a region* T *is independent of the particular choice of Cartesian coordinates. For any* P *in* T *it is given by* (11).

Equation (11) is sometimes used as a *definition* of the divergence. Then the representation (1) in Cartesian coordinates can be derived from (11).

Further applications of the divergence theorem follow in the problem set and in the next section. The examples in the next section will also shed further light on the nature of the divergence.

PROBLEM SET 10.7

1–8 APPLICATION: MASS DISTRIBUTION

Find the total mass of a mass distribution of density σ in a region T in space.

1. $\sigma = x^2 + y^2 + z^2$, T the box $|x| \leqq 4$, $|y| \leqq 1$, $0 \leqq z \leqq 2$
2. $\sigma = xyz$, T the box $0 \leqq x \leqq a$, $0 \leqq y \leqq b$, $0 \leqq z \leqq c$
3. $\sigma = e^{-x-y-z}$, T: $0 \leqq x \leqq 1 - y$, $0 \leqq y \leqq 1$, $0 \leqq z \leqq 2$
4. σ as in Prob. 3, T the tetrahedron with vertices (0, 0, 0), (3, 0, 0), (0, 3, 0), (0, 0, 3)
5. $\sigma = \sin 2x \cos 2y$, T: $0 \leqq x \leqq \frac{1}{4}\pi$, $\frac{1}{4}\pi - x \leqq y \leqq \frac{1}{4}\pi$, $0 \leqq z \leqq 6$
6. $\sigma = x^2y^2z^2$, T the cylindrical region $x^2 + z^2 \leqq 16$, $|y| \leqq 4$
7. $\sigma = \arctan(y/x)$, T: $x^2 + y^2 + z^2 \leqq a^2$, $z \geqq 0$
8. $\sigma = x^2 + y^2$, T as in Prob. 7

9–18 APPLICATION OF THE DIVERGENCE THEOREM

Evaluate the surface integral $\iint_S \mathbf{F} \cdot \mathbf{n}\, dA$ by the divergence theorem. Show the details.

9. $\mathbf{F} = [x^2, 0, z^2]$, S the surface of the box $|x| \leqq 1$, $|y| \leqq 3$, $0 \leqq z \leqq 2$
10. Solve Prob. 9 by direct integration.
11. $\mathbf{F} = [e^x, e^y, e^z]$, S the surface of the cube $|x| \leqq 1$, $|y| \leqq 1$, $|z| \leqq 1$
12. $\mathbf{F} = [x^3 - y^3, y^3 - z^3, z^3 - x^3]$, S the surface of $x^2 + y^2 + z^2 \leqq 25$, $z \geqq 0$
13. $\mathbf{F} = [\sin y, \cos x, \cos z]$, S, the surface of $x^2 + y^2 \leqq 4$, $|z| \leqq 2$ (a cylinder and two disks!)
14. $\mathbf{F}$ as in Prob. 13, S the surface of $x^2 + y^2 \leqq 9$, $0 \leqq z \leqq 2$

15. $\mathbf{F} = [2x^2, \frac{1}{2}y^2, \sin \pi z]$, S the surface of the tetrahedron with vertices (0, 0, 0), (1, 0, 0), (0, 1, 0), (0, 0, 1)

16. $\mathbf{F} = [\cosh x, z, y]$, S as in Prob. 15

17. $\mathbf{F} = [x^2, y^2, z^2]$, S the surface of the cone $x^2 + y^2 \leqq z^2$, $0 \leqq z \leqq h$

18. $\mathbf{F} = [xy, yz, zx]$, S the surface of the cone $x^2 + y^2 \leqq 4z^2$, $0 \leqq z \leqq 2$

19–23 APPLICATION: MOMENT OF INERTIA

Given a mass of density 1 in a region T of space, find the moment of intertia about the x-axis

$$I_x = \iiint_T (y^2 + z^2)\, dx\, dy\, dz.$$

19. The box $-a \leqq x \leqq a, \quad -b \leqq y \leqq b, \quad -c \leqq z \leqq c$

20. The ball $x^2 + y^2 + z^2 \leqq a^2$

21. The cylinder $y^2 + z^2 \leqq a^2, \quad 0 \leqq x \leqq h$

22. The paraboloid $y^2 + z^2 \leqq x, \quad 0 \leqq x \leqq h$

23. The cone $y^2 + z^2 \leqq x^2, \quad 0 \leqq x \leqq h$

24. Why is I_x in Prob. 23 for large h larger than I_x in Prob. 22 (and the same h)? Why is it smaller for $h = 1$? Give physical reason.

25. Show that for a solid of revolution, $I_x = \dfrac{\pi}{2}\displaystyle\int_0^h r^4(x)\, dx.$ Solve Probs. 20–23 by this formula.

10.8 Further Applications of the Divergence Theorem

The divergence theorem has many important applications: In *fluid flow*, it helps characterize *sources* and *sinks* of fluids. In *heat flow*, it leads to the *heat equation.* In *potential theory*, it gives properties of the solutions of *Laplace's equation.* In this section, we assume that the region T and its boundary surface S are such that the divergence theorem applies.

EXAMPLE 1 **Fluid Flow. Physical Interpretation of the Divergence**

From the divergence theorem we may obtain an intuitive interpretation of the divergence of a vector. For this purpose we consider the flow of an incompressible fluid (see Sec. 9.8) of constant density $\rho = 1$ which is **steady**, that is, does not vary with time. Such a flow is determined by the field of its velocity vector $\mathbf{v}(P)$ at any point P.

Let S be the boundary surface of a region T in space, and let $\mathbf{n}$ be the outer unit normal vector of S. Then $\mathbf{v} \bullet \mathbf{n}$ is the normal component of $\mathbf{v}$ in the direction of $\mathbf{n}$, and $|\mathbf{v} \bullet \mathbf{n}\, dA|$ is the mass of fluid *leaving* T (if $\mathbf{v} \bullet \mathbf{n} > 0$ at some P) or *entering* T (if $\mathbf{v} \bullet \mathbf{n} < 0$ at P) per unit time at some point P of S through a small portion ΔS of S of area ΔA. Hence the total mass of fluid that flows across S from T to the outside per unit time is given by the surface integral

$$\iint_S \mathbf{v} \bullet \mathbf{n}\, dA.$$

Division by the volume V of T gives the **average flow** out of T:

$$\frac{1}{V}\iint_S \mathbf{v} \bullet \mathbf{n}\, dA. \tag{1}$$

Since the flow is steady and the fluid is incompressible, the amount of fluid flowing outward must be continuously supplied. Hence, if the value of the integral (1) is different from zero, there must be **sources** (*positive sources and negative sources, called* **sinks**) in T, that is, points where fluid is produced or disappears.

If we let T shrink down to a fixed point P in T, we obtain from (1) the **source intensity** at P given by the right side of (11) in the last section with $\mathbf{F} \bullet \mathbf{n}$ replaced by $\mathbf{v} \bullet \mathbf{n}$, that is,

$$\operatorname{div} \mathbf{v}(P) = \lim_{d(T)\to 0} \frac{1}{V(T)} \iint_{S(T)} \mathbf{v} \bullet \mathbf{n}\, dA. \tag{2}$$

Hence *the divergence of the velocity vector* **v** *of a steady incompressible flow is the source intensity of the flow at the corresponding point.*

There are no sources in T if and only if div **v** is zero everywhere in T. Then for any closed surface S in T we have

$$\iint_S \mathbf{v} \bullet \mathbf{n}\, dA = 0.$$

EXAMPLE 2 Modeling of Heat Flow. Heat or Diffusion Equation

Physical experiments show that in a body, heat flows in the direction of decreasing temperature, and the rate of flow is proportional to the gradient of the temperature. This means that the velocity **v** of the heat flow in a body is of the form

$$\mathbf{v} = -K \operatorname{grad} U \tag{3}$$

where $U(x, y, z, t)$ is temperature, t is time, and K is called the *thermal conductivity* of the body; in ordinary physical circumstances K is a constant. Using this information, set up the mathematical model of heat flow, the so-called **heat equation** or **diffusion equation**.

Solution. Let T be a region in the body bounded by a surface S with outer unit normal vector **n** such that the divergence theorem applies. Then $\mathbf{v} \cdot \mathbf{n}$ is the component of **v** in the direction of **n**, and the amount of heat leaving T per unit time is

$$\iint_S \mathbf{v} \bullet \mathbf{n}\, dA.$$

This expression is obtained similarly to the corresponding surface integral in the last example. Using

$$\operatorname{div}(\operatorname{grad} U) = \nabla^2 U = U_{xx} + U_{yy} + U_{zz}$$

(the Laplacian; see (3) in Sec. 9.8), we have by the divergence theorem and (3)

$$\begin{aligned} \iint_S \mathbf{v} \bullet \mathbf{n}\, dA &= -K \iiint_T \operatorname{div}(\operatorname{grad} U)\, dx\, dy\, dz \\ &= -K \iiint_T \nabla^2 U\, dx\, dy\, dz. \end{aligned} \tag{4}$$

On the other hand, the total amount of heat H in T is

$$H = \iiint_T \sigma \rho U\, dx\, dy\, dz$$

where the constant σ is the specific heat of the material of the body and ρ is the density (= mass per unit volume) of the material. Hence the time rate of decrease of H is

$$-\frac{\partial H}{\partial t} = -\iiint_T \sigma \rho \frac{\partial U}{\partial t}\, dx\, dy\, dz$$

and this must be equal to the above amount of heat leaving T. From (4) we thus have

$$-\iiint_T \sigma \rho \frac{\partial U}{\partial t}\, dx\, dy\, dz = -K \iiint_T \nabla^2 U\, dx\, dy\, dz$$

or

$$\iiint_T \left(\sigma \rho \frac{\partial U}{\partial t} - K \nabla^2 U\right) dx\, dy\, dz = 0.$$

Since this holds for any region T in the body, the integrand (if continuous) must be zero everywhere; that is,

$$\frac{\partial U}{\partial t} = c^2 \nabla^2 U \qquad c^2 = \frac{K}{\sigma \rho} \tag{5}$$

where c^2 is called the *thermal diffusivity* of the material. This partial differential equation is called the **heat equation**. It is the fundamental equation for heat conduction. And our derivation is another impressive demonstration of the great importance of the divergence theorem. Methods for solving heat problems will be shown in Chap. 12.

The heat equation is also called the **diffusion equation** because it also models diffusion processes of motions of molecules tending to level off differences in density or pressure in gases or liquids.

If heat flow does not depend on time, it is called **steady-state heat flow**. Then $\partial U/\partial t = 0$, so that (5) reduces to *Laplace's equation* $\nabla^2 U = 0$. We met this equation in Secs. 9.7 and 9.8, and we shall now see that the divergence theorem adds basic insights into the nature of solutions of this equation. ■

Potential Theory. Harmonic Functions

The theory of solutions of **Laplace's equation**

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} + \frac{\partial^2 f}{\partial z^2} = 0 \tag{6}$$

is called **potential theory**. A solution of (6) with ***continuous*** second-order partial derivatives is called a **harmonic function**. That continuity is needed for application of the divergence theorem in potential theory, where the theorem plays a key role that we want to explore. Further details of potential theory follow in Chaps. 12 and 18.

EXAMPLE 3 A Basic Property of Solutions of Laplace's Equation

The integrands in the divergence theorem are div $\mathbf{F}$ and $\mathbf{F} \cdot \mathbf{n}$ (Sec. 10.7). If $\mathbf{F}$ is the gradient of a scalar function, say, $\mathbf{F} = \operatorname{grad} f$, then $\operatorname{div} \mathbf{F} = \operatorname{div}(\operatorname{grad} f) = \nabla^2 f$; see (3), Sec. 9.8. Also, $\mathbf{F} \cdot \mathbf{n} = \mathbf{n} \cdot \mathbf{F} = \mathbf{n} \cdot \operatorname{grad} f$. This is the directional derivative of f in the outer normal direction of S, the boundary surface of the region T in the theorem. This derivative is called the (outer) **normal derivative** of f and is denoted by $\partial f/\partial n$. Thus the formula in the divergence theorem becomes

$$\iiint_T \nabla^2 f \, dV = \iint_S \frac{\partial f}{\partial n} \, dA. \tag{7}$$

This is the three-dimensional analog of (9) in Sec. 10.4. Because of the assumptions in the divergence theorem this gives the following result. ■

THEOREM 1 A Basic Property of Harmonic Functions

Let $f(x, y, z)$ be a harmonic function in some domain D is space. Let S be any piecewise smooth closed orientable surface in D whose entire region it encloses belongs to D. Then the integral of the normal derivative of f taken over S is zero. (For "piecewise smooth" see Sec. 10.5.)

EXAMPLE 4 Green's Theorems

Let f and g be scalar functions such that $F = f \operatorname{grad} g$ satisfies the assumptions of the divergence theorem in some region T. Then

$$\begin{aligned}\operatorname{div} \mathbf{F} &= \operatorname{div}(f \operatorname{grad} g)\\ &= \operatorname{div}\left(\left[f\frac{\partial g}{\partial x}, f\frac{\partial g}{\partial y}, f\frac{\partial g}{\partial z}\right]\right)\\ &= \left(\frac{\partial f}{\partial x}\frac{\partial g}{\partial x} + f\frac{\partial^2 g}{\partial x^2}\right) + \left(\frac{\partial f}{\partial y}\frac{\partial g}{\partial y} + f\frac{\partial^2 g}{\partial y^2}\right) + \left(\frac{\partial f}{\partial z}\frac{\partial g}{\partial z} + f\frac{\partial^2 g}{\partial z^2}\right)\\ &= f\nabla^2 g + \operatorname{grad} f \bullet \operatorname{grad} g.\end{aligned}$$

Also, since f is a scalar function,

$$\begin{aligned}\mathbf{F} \bullet \mathbf{n} &= \mathbf{n} \bullet \mathbf{F}\\ &= \mathbf{n} \bullet (f \operatorname{grad} g)\\ &= (\mathbf{n} \bullet \operatorname{grad} g) f.\end{aligned}$$

Now $\mathbf{n} \bullet \operatorname{grad} g$ is the directional derivative $\partial g/\partial n$ of g in the outer normal direction of S. Hence the formula in the divergence theorem becomes "**Green's first formula**"

(8)
$$\iiint_T (f\nabla^2 g + \operatorname{grad} f \bullet \operatorname{grad} g)\, dV = \iint_S f \frac{\partial g}{\partial n}\, dA.$$

Formula (8) together with the assumptions is known as the *first form of Green's theorem.*

Interchanging f and g we obtain a similar formula. Subtracting this formula from (8) we find

(9)
$$\iiint_T (f\nabla^2 g - g\nabla^2 f)\, dV = \iint_S \left(f\frac{\partial g}{\partial n} - g\frac{\partial f}{\partial n}\right) dA.$$

This formula is called **Green's second formula** or (together with the assumptions) the *second form of Green's theorem.* ■

EXAMPLE 5 Uniqueness of Solutions of Laplace's Equation

Let f be harmonic in a domain D and let f be zero everywhere on a piecewise smooth closed orientable surface S in D whose entire region T it encloses belongs to D. Then $\nabla^2 g$ is zero in T, and the surface integral in (8) is zero, so that (8) with $g = f$ gives

$$\iiint_T \operatorname{grad} f \bullet \operatorname{grad} f\, dV = \iiint_T |\operatorname{grad} f|^2\, dV = 0.$$

Since f is harmonic, $\operatorname{grad} f$ and thus $|\operatorname{grad} f|$ are continuous in T and on S, and since $|\operatorname{grad} f|$ is nonnegative, to make the integral over T zero, $\operatorname{grad} f$ must be the zero vector everywhere in T. Hence $f_x = f_y = f_z = 0$, and f is constant in T and, because of continuity, it is equal to its value 0 on S. This proves the following theorem.

THEOREM 2

Harmonic Functions

Let $f(x, y, z)$ be harmonic in some domain D and zero at every point of a piecewise smooth closed orientable surface S in D whose entire region T it encloses belongs to D. Then f is identically zero in T.

This theorem has an important consequence. Let f_1 and f_2 be functions that satisfy the assumptions of Theorem 1 and take on the same values on S. Then their difference $f_1 - f_2$ satisfies those assumptions and has the value 0 everywhere on S. Hence, Theorem 2 implies that

$$f_1 - f_2 = 0 \qquad \text{throughout} \qquad T,$$

and we have the following fundamental result.

THEOREM 3

Uniqueness Theorem for Laplace's Equation

Let T be a region that satisfies the assumptions of the divergence theorem, and let $f(x, y, z)$ be a harmonic function in a domain D that contains T and its boundary surface S. Then f is uniquely determined in T by its values on S.

The problem of determining a solution u of a partial differential equation in a region T such that u assumes given values on the boundary surface S of T is called the **Dirichlet problem**.[8] We may thus reformulate Theorem 3 as follows.

THEOREM 3*

Uniqueness Theorem for the Dirichlet Problem

If the assumptions in Theorem 3 are satisfied and the Dirichlet problem for the Laplace equation has a solution in T, then this solution is unique.

These theorems demonstrate the extreme importance of the divergence theorem in potential theory. ■

PROBLEM SET 10.8

1–6 VERIFICATIONS

1. **Harmonic functions.** Verify Theorem 1 for $f = 2z^2 - x^2 - y^2$ and S the surface of the box $0 \leqq x \leqq a$, $0 \leqq y \leqq b$, $0 \leqq z \leqq c$.
2. **Harmonic functions.** Verify Theorem 1 for $f = x^2 - y^2$ and the surface of the cylinder $x^2 + y^2 = 4$, $0 \leqq z \leqq h$.
3. **Green's first identity.** Verify (8) for $f = 4y^2$, $g = x^2$, S the surface of the "unit cube" $0 \leqq x \leqq 1$, $0 \leqq y \leqq 1$, $0 \leqq z \leqq 1$. What are the assumptions on f and g in (8)? Must f and g be harmonic?
4. **Green's first identity.** Verify (8) for $f = x$, $g = y^2 + z^2$, S the surface of the box $0 \leqq x \leqq 1$, $0 \leqq y \leqq 2$, $0 \leqq z \leqq 3$.

[8] PETER GUSTAV LEJEUNE DIRICHLET (1805–1859), German mathematician, studied in Paris under Cauchy and others and succeeded Gauss at Göttingen in 1855. He became known by his important research on Fourier series (he knew Fourier personally) and in number theory.

5. Green's second identity. Verify (9) for $f = 6y^2$, $g = 2x^2$, S the unit cube in Prob. 3.

6. Green's second identity. Verify (9) for $f = x^2$, $g = y^4$, S the unit cube in Prob. 3.

7–11 VOLUME

Use the divergence theorem, assuming that the assumptions on T and S are satisfied.

7. Show that a region T with boundary surface S has the volume

$$V = \iint_S x\,dy\,dz = \iint_S y\,dz\,dx = \iint_S z\,dx\,dy$$
$$= \frac{1}{3}\iint_S (x\,dy\,dz + y\,dz\,dx + z\,dx\,dy).$$

8. Cone. Using the third expression for v in Prob. 7, verify $V = \pi a^2 h/3$ for the volume of a circular cone of height h and radius of base a.

9. Ball. Find the volume under a hemisphere of radius a from in Prob. 7.

10. Volume. Show that a region T with boundary surface S has the volume

$$V = \frac{1}{3}\iint_S r\cos\phi\,dA$$

where r is the distance of a variable point $P{:}\,(x, y, z)$ on S from the origin O and ϕ is the angle between the directed line OP and the outer normal of S at P. Make a sketch. *Hint.* Use (2) in Sec. 10.7 with $\mathbf{F} = [x, y, z]$.

11. Ball. Find the volume of a ball of radius a from Prob. 10.

12. TEAM PROJECT. Divergence Theorem and Potential Theory. The importance of the divergence theorem in potential theory is obvious from (7)–(9) and Theorems 1–3. To emphasize it further, consider functions f and g that are harmonic in some domain D containing a region T with boundary surface S such that T satisfies the assumptions in the divergence theorem. Prove, and illustrate by examples, that then:

(a) $$\iint_S g\frac{\partial g}{\partial n}\,dA = \iiint_T |\text{grad } g|^2\,dV.$$

(b) If $\partial g/\partial n = 0$ on S, then g is constant in T.

(c) $$\iint_S \left(f\frac{\partial g}{\partial n} - g\frac{\partial f}{\partial n}\right)dA = 0.$$

(d) If $\partial f/\partial n = \partial g/\partial n$ on S, then $f = g + c$ in T, where c is a constant.

(e) The **Laplacian** can be represented ***independently of coordinate systems*** in the form

$$\nabla^2 f = \lim_{d(T)\to 0}\frac{1}{V(T)}\iint_{S(T)}\frac{\partial f}{\partial n}\,dA$$

where $d(T)$ is the maximum distance of the points of a region T bounded by $S(T)$ from the point at which the Laplacian is evaluated and $V(T)$ is the volume of T.

10.9 Stokes's Theorem

Let us review some of the material covered so far. Double integrals over a region in the plane can be transformed into line integrals over the boundary curve of that region and conversely, line integrals into double integrals. This important result is known as *Green's theorem in the plane* and was explained in Sec. 10.4. We also learned that we can transform triple integrals into surface integrals and vice versa, that is, surface integrals into triple integrals. This "big" theorem is called *Gauss's divergence theorem* and was shown in Sec. 10.7.

To complete our discussion on transforming integrals, we now introduce another "big" theorem that allows us to transform surface integrals into line integrals and conversely, line integrals into surface integrals. It is called **Stokes's Theorem**, and it generalizes Green's theorem in the plane (see Example 2 below for this immediate observation). Recall from Sec. 9.9 that

$$\text{curl}\,\mathbf{F} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \partial/\partial x & \partial/\partial y & \partial/\partial z \\ F_1 & F_2 & F_3 \end{vmatrix} \tag{1}$$

which we will need immediately.

THEOREM 1

Stokes's Theorem[9]
(Transformation Between Surface and Line Integrals)

Let S be a piecewise smooth[9] oriented surface in space and let the boundary of S be a piecewise smooth simple closed curve C. Let $\mathbf{F}(x, y, z)$ *be a continuous vector function that has continuous first partial derivatives in a domain in space containing S. Then*

$$\iint_S (\text{curl } \mathbf{F}) \cdot \mathbf{n} \, dA = \oint_C \mathbf{F} \cdot \mathbf{r}'(s) \, ds. \tag{2}$$

Here $\mathbf{n}$ *is a unit normal vector of S and, depending on* $\mathbf{n}$*, the integration around C is taken in the sense shown in Fig.* 254*. Furthermore,* $\mathbf{r}' = d\mathbf{r}/ds$ *is the unit tangent vector and s the arc length of C.*

In components, formula (2) *becomes*

$$\begin{aligned} &\iint_R \left[\left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}\right)N_1 + \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}\right)N_2 + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right)N_3\right] du\, dv \\ &\qquad = \oint_{\bar{C}} (F_1\, dx + F_2\, dy + F_3\, dz). \end{aligned} \tag{2*}$$

Here, $\mathbf{F} = [F_1, \ F_2, \ F_3]$, $\mathbf{N} = [N_1, \ N_2, \ N_3]$, $\mathbf{n}\, dA = \mathbf{N}\, du\, dv$, $\mathbf{r}'\, ds = [dx, \ dy, \ dz]$, *and R is the region with boundary curve* $\bar{C}$ *in the uv-plane corresponding to S represented by* $\mathbf{r}(u, v)$.

The proof follows after Example 1.

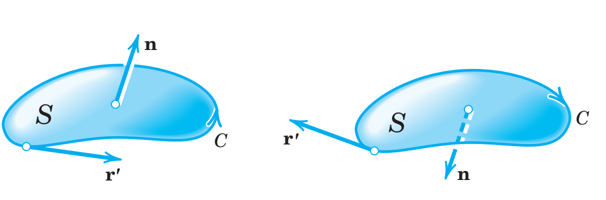


Fig. 254. Stokes's theorem

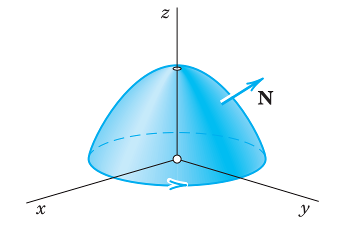


Fig. 255. Surface *S* in Example 1

EXAMPLE 1

Verification of Stokes's Theorem

Before we prove Stokes's theorem, let us first get used to it by verifying it for $\mathbf{F} = [y, z, x]$ and S the paraboloid (Fig. 255)

$$z = f(x, y) = 1 - (x^2 + y^2), \qquad z \geqq 0.$$

---

[9]Sir GEORGE GABRIEL STOKES (1819–1903), Irish mathematician and physicist who became a professor in Cambridge in 1849. He is also known for his important contribution to the theory of infinite series and to viscous flow (Navier–Stokes equations), geodesy, and optics.

"Piecewise smooth" curves and surfaces are defined in Secs. 10.1 and 10.5.

Solution. The curve C, oriented as in Fig. 255, is the circle $\mathbf{r}(s) = [\cos s, \sin s, 0]$. Its unit tangent vector is $\mathbf{r}'(s) = [-\sin s, \cos s, 0]$. The function $\mathbf{F} = [y, z, x]$ on C is $\mathbf{F}(\mathbf{r}(s)) = [\sin s, 0, \cos s]$. Hence

$$\oint_C \mathbf{F} \bullet d\mathbf{r} = \int_0^{2\pi} \mathbf{F}(\mathbf{r}(s)) \bullet \mathbf{r}'(s)\, ds = \int_0^{2\pi} [(\sin s)(-\sin s) + 0 + 0]\, ds = -\pi.$$

We now consider the surface integral. We have $F_1 = y$, $F_2 = z$, $F_3 = x$, so that in (2*) we obtain

$$\operatorname{curl} \mathbf{F} = \operatorname{curl}\,[F_1, \quad F_2, \quad F_3] = \operatorname{curl}\,[y, \quad z, \quad x] = [-1, \quad -1, \quad -1].$$

A normal vector of S is $\mathbf{N} = \operatorname{grad}(z - f(x, y)) = [2x, 2y, 1]$. Hence $(\operatorname{curl} \mathbf{F}) \bullet \mathbf{N} = -2x - 2y - 1$. Now $\mathbf{n}\, dA = \mathbf{N}\, dx\, dy$ (see (3*) in Sec. 10.6 with x, y instead of u, v). Using polar coordinates r, θ defined by $x = r\cos\theta$, $y = r\sin\theta$ and denoting the projection of S into the xy-plane by R, we thus obtain

$$\iint_S (\operatorname{curl} \mathbf{F}) \bullet \mathbf{n}\, dA = \iint_R (\operatorname{curl} \mathbf{F}) \bullet \mathbf{N}\, dx\, dy = \iint_R (-2x - 2y - 1)\, dx\, dy$$

$$= \int_{\theta=0}^{2\pi} \int_{r=0}^{1} (-2r(\cos\theta + \sin\theta) - 1) r\, dr\, d\theta$$

$$= \int_{\theta=0}^{2\pi} \left(-\frac{2}{3}(\cos\theta + \sin\theta) - \frac{1}{2}\right) d\theta = 0 + 0 - \frac{1}{2}(2\pi) = -\pi. \quad \blacksquare$$

PROOF We prove Stokes's theorem. Obviously, (2) holds if the integrals of each component on both sides of (2*) are equal; that is,

$$\text{(3)} \qquad \iint_R \left(\frac{\partial F_1}{\partial z} N_2 - \frac{\partial F_1}{\partial y} N_3\right) du\, dv = \oint_{\bar{C}} F_1\, dx$$

$$\text{(4)} \qquad \iint_R \left(-\frac{\partial F_2}{\partial z} N_1 + \frac{\partial F_2}{\partial x} N_3\right) du\, dv = \oint_{\bar{C}} F_2\, dy$$

$$\text{(5)} \qquad \iint_R \left(\frac{\partial F_3}{\partial y} N_1 - \frac{\partial F_3}{\partial x} N_2\right) du\, dv = \oint_{\bar{C}} F_3\, dz.$$

We prove this first for a surface S that can be represented simultaneously in the forms

(6) (a) $z = f(x, y)$, (b) $y = g(x, z)$, (c) $x = h(y, z)$.

We prove (3), using (6a). Setting $u = x$, $v = y$, we have from (6a)

$$\mathbf{r}(u, v) = \mathbf{r}(x, y) = [x, y, f(x, y)] = x\mathbf{i} + y\mathbf{j} + f\mathbf{k}$$

and in (2), Sec. 10.6, by direct calculation

$$\mathbf{N} = \mathbf{r}_u \times \mathbf{r}_v = \mathbf{r}_x \times \mathbf{r}_y = [-f_x, \quad -f_y, \quad 1] = -f_x\mathbf{i} - f_y\mathbf{j} + \mathbf{k}.$$

Note that $\mathbf{N}$ is an *upper* normal vector of S, since it has a *positive* z-component. Also, $R = S^*$, the projection of S into the xy-plane, with boundary curve $\bar{C} = C^*$ (Fig. 256). Hence the left side of (3) is

$$\iint_{S^*} \left[\frac{\partial F_1}{\partial z}(-f_y) - \frac{\partial F_1}{\partial y} \right] dx\, dy. \tag{7}$$

We now consider the right side of (3). We transform this line integral over $\bar{C} = C^*$ into a double integral over S^* by applying Green's theorem [formula (1) in Sec. 10.4 with $F_2 = 0$]. This gives

$$\oint_{C^*} F_1\, dx = \iint_{S^*} -\frac{\partial F_1}{\partial y}\, dx\, dy.$$

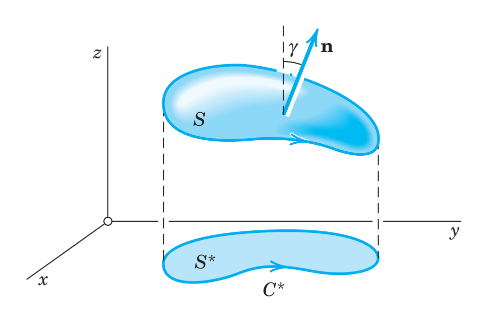


Fig. 256. Proof of Stokes's theorem

Here, $F_1 = F_1(x, y, f(x, y))$. Hence by the chain rule (see also Prob. 10 in Problem Set 9.6),

$$-\frac{\partial F_1(x, y, f(x, y))}{\partial y} = -\frac{\partial F_1(x, y, z)}{\partial y} - \frac{\partial F_1(x, y, z)}{\partial z}\frac{\partial f}{\partial y} \qquad [z = f(x, y)].$$

We see that the right side of this equals the integrand in (7). This proves (3). Relations (4) and (5) follow in the same way if we use (6b) and (6c), respectively. By addition we obtain (2*). This proves Stokes's theorem for a surface S that can be represented simultaneously in the forms (6a), (6b), (6c).

As in the proof of the divergence theorem, our result may be immediately extended to a surface S that can be decomposed into finitely many pieces, each of which is of the kind just considered. This covers most of the cases of practical interest. The proof in the case of a most general surface S satisfying the assumptions of the theorem would require a limit process; this is similar to the situation in the case of Green's theorem in Sec. 10.4. ■

EXAMPLE 2 Green's Theorem in the Plane as a Special Case of Stokes's Theorem

Let $\mathbf{F} = [F_1, F_2] = F_1\mathbf{i} + F_2\mathbf{j}$ be a vector function that is continuously differentiable in a domain in the xy-plane containing a simply connected bounded closed region S whose boundary C is a piecewise smooth simple closed curve. Then, according to (1),

$$(\text{curl } \mathbf{F}) \bullet \mathbf{n} = (\text{curl } \mathbf{F}) \bullet \mathbf{k} = \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}.$$

Hence the formula in Stokes's theorem now takes the form

$$\iint_S \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dA = \oint_C (F_1\,dx + F_2\,dy).$$

This shows that Green's theorem in the plane (Sec. 10.4) is a special case of Stokes's theorem (which we needed in the proof of the latter!). ■

EXAMPLE 3 Evaluation of a Line Integral by Stokes's Theorem

Evaluate $\int_C \mathbf{F} \cdot \mathbf{r}'\,ds$, where C is the circle $x^2 + y^2 = 4$, $z = -3$, oriented counterclockwise as seen by a person standing at the origin, and, with respect to right-handed Cartesian coordinates,

$$\mathbf{F} = [y, \quad xz^3, \quad -zy^3] = y\mathbf{i} + xz^3\mathbf{j} - zy^3\mathbf{k}.$$

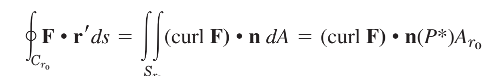
Solution. As a surface S bounded by C we can take the plane circular disk $x^2 + y^2 \leqq 4$ in the plane $z = -3$. Then $\mathbf{n}$ in Stokes's theorem points in the positive z-direction; thus $\mathbf{n} = \mathbf{k}$. Hence $(\text{curl } \mathbf{F}) \cdot \mathbf{n}$ is simply the component of curl $\mathbf{F}$ in the positive z-direction. Since $\mathbf{F}$ with $z = -3$ has the components $F_1 = y$, $F_2 = -27x$, $F_3 = 3y^3$, we thus obtain

$$(\text{curl } \mathbf{F}) \cdot \mathbf{n} = \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} = -27 - 1 = -28.$$

Hence the integral over S in Stokes's theorem equals -28 times the area 4π of the disk S. This yields the answer $-28 \cdot 4\pi = -112\pi \approx -352$. Confirm this by direct calculation, which involves somewhat more work. ■

EXAMPLE 4 Physical Meaning of the Curl in Fluid Motion. Circulation

Let S_{r_0} be a circular disk of radius r_0 and center P bounded by the circle C_{r_0} (Fig. 257), and let $\mathbf{F}(Q) \equiv \mathbf{F}(x, y, z)$ be a continuously differentiable vector function in a domain containing S_{r_0}. Then by Stokes's theorem and the mean value theorem for surface integrals (see Sec. 10.6),



$$\oint_{C_{r_0}} \mathbf{F} \cdot \mathbf{r}'\,ds = \iint_{S_{r_0}} (\text{curl } \mathbf{F}) \cdot \mathbf{n}\,dA = (\text{curl } \mathbf{F}) \cdot \mathbf{n}(P^*)A_{r_0}$$

Fig. 257. Example 4

where A_{r_0} is the area of S_{r_0} and P^* is a suitable point of S_{r_0}. This may be written in the form

$$(\text{curl } \mathbf{F}) \cdot \mathbf{n}(P^*) = \frac{1}{A_{r_0}} \oint_{C_{r_0}} \mathbf{F} \cdot \mathbf{r}'\,ds.$$

In the case of a fluid motion with velocity vector $\mathbf{F} = \mathbf{v}$, the integral

$$\oint_{C_{r_0}} \mathbf{v} \cdot \mathbf{r}'\,ds$$

is called the **circulation** of the flow around C_{r_0}. It measures the extent to which the corresponding fluid motion is a rotation around the circle C_{r_0}. If we now let r_0 approach zero, we find

$$(8) \qquad (\text{curl } \mathbf{v}) \cdot \mathbf{n}(P) = \lim_{r_0 \to 0} \frac{1}{A_{r_0}} \oint_{C_{r_0}} \mathbf{v} \cdot \mathbf{r}'\,ds;$$

that is, the component of the curl in the positive normal direction can be regarded as the **specific circulation** (circulation per unit area) of the flow in the surface at the corresponding point. ■

EXAMPLE 5 Work Done in the Displacement around a Closed Curve

Find the work done by the force $\mathbf{F} = 2xy^3 \sin z\,\mathbf{i} + 3x^2y^2 \sin z\,\mathbf{j} + x^2y^3 \cos z\,\mathbf{k}$ in the displacement around the curve of intersection of the paraboloid $z = x^2 + y^2$ and the cylinder $(x - 1)^2 + y^2 = 1$.

Solution. This work is given by the line integral in Stokes's theorem. Now $\mathbf{F} = \text{grad } f$, where $f = x^2y^3 \sin z$ and $\text{curl}(\text{grad } f) = \mathbf{0}$ (see (2) in Sec. 9.9), so that $(\text{curl } \mathbf{F}) \cdot \mathbf{n} = 0$ and the work is 0 by Stokes's theorem. This agrees with the fact that the present field is conservative (definition in Sec. 9.7). ■

Stokes's Theorem Applied to Path Independence

We emphasized in Sec. 10.2 that the value of a line integral generally depends not only on the function to be integrated and on the two endpoints A and B of the path of integration C, but also on the particular choice of a path from A to B. In Theorem 3 of Sec. 10.2 we proved that if a line integral

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) \tag{9}$$

(involving continuous F_1, F_2, F_3 that have continuous first partial derivatives) is path independent in a domain D, then curl $\mathbf{F} = \mathbf{0}$ in D. And we claimed in Sec. 10.2 that, conversely, curl $\mathbf{F} = \mathbf{0}$ everywhere in D implies path independence of (9) in D provided D is *simply connected.* A proof of this needs Stokes's theorem and can now be given as follows.

Let C be any closed path in D. Since D is simply connected, we can find a surface S in D bounded by C. Stokes's theorem applies and gives

$$\oint_C (F_1\,dx + F_2\,dy + F_3\,dz) = \oint_C \mathbf{F} \cdot \mathbf{r}'\,ds = \iint_S (\text{curl}\,\mathbf{F}) \cdot \mathbf{n}\,dA$$

for proper direction on C and normal vector $\mathbf{n}$ on S. Since curl $\mathbf{F} = \mathbf{0}$ in D, the surface integral and hence the line integral are zero. This and Theorem 2 of Sec. 10.2 imply that the integral (9) is path independent in D. This completes the proof. ■

PROBLEM SET 10.9

1–10 DIRECT INTEGRATION OF SURFACE INTEGRALS

Evaluate the surface integral $\iint_S (\text{curl}\,\mathbf{F}) \cdot \mathbf{n}\,dA$ directly for the given $\mathbf{F}$ and S.

1. $\mathbf{F} = [z^2, -x^2, 0]$, S the rectangle with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(0, 4, 4)$, $(1, 4, 4)$
2. $\mathbf{F} = [-13 \sin y, 3 \sinh z, x]$, S the rectangle with vertices $(0, 0, 2)$, $(4, 0, 2)$, $(4, \pi/2, 2)$, $(0, \pi/2, 2)$
3. $\mathbf{F} = [e^{-z}, e^{-z} \cos y, e^{-z} \sin y]$, S: $z = y^2/2$, $-1 \leqq x \leqq 1$, $0 \leqq y \leqq 1$
4. $\mathbf{F}$ as in Prob. 1, $z = xy$ $(0 \leqq x \leqq 1,\ 0 \leqq y \leqq 4)$. Compare with Prob. 1.
5. $\mathbf{F} = [z^2, \frac{3}{2}x, 0]$, S: $0 \leqq x \leqq a$, $0 \leqq y \leqq a$, $z = 1$
6. $\mathbf{F} = [y^3, -x^3, 0]$, S: $x^2 + y^2 \leqq 1$, $z = 0$
7. $\mathbf{F} = [e^y, e^z, e^x]$, S: $z = x^2$ $(0 \leqq x \leqq 2,\ 0 \leqq y \leqq 1)$
8. $\mathbf{F} = [z^2, x^2, y^2]$, S: $z = \sqrt{x^2 + y^2}$, $y \geqq 0$, $0 \leqq z \leqq h$
9. Verify Stokes's theorem for $\mathbf{F}$ and S in Prob. 5.
10. Verify Stokes's theorem for $\mathbf{F}$ and S in Prob. 6.
11. **Stokes's theorem not applicable.** Evaluate $\oint_C \mathbf{F} \cdot \mathbf{r}'\,ds$, $\mathbf{F} = (x^2 + y^2)^{-1}[-y, x]$, C: $x^2 + y^2 = 1$, $z = 0$, oriented clockwise. Why can Stokes's theorem not be applied? What (false) result would it give?
12. **WRITING PROJECT. Grad, Div, Curl in Connection with Integrals.** Make a list of ideas and results on this topic in this chapter. See whether you can rearrange or combine parts of your material. Then subdivide the material into 3–5 portions and work out the details of each portion. Include no proofs but simple typical examples of your own that lead to a better understanding of the material.

13–20 EVALUATION OF $\oint_C \mathbf{F} \cdot \mathbf{r}'\,ds$

Calculate this line integral by Stokes's theorem for the given $\mathbf{F}$ and C. Assume the Cartesian coordinates to be right-handed and the z-component of the surface normal to be nonnegative.

13. $\mathbf{F} = [-5y, 4x, z]$, C the circle $x^2 + y^2 = 16$, $z = 4$

14. $\mathbf{F} = [z^3, x^3, y^3]$, C the circle $x = 2$, $y^2 + z^2 = 9$

15. $\mathbf{F} = [y^2, x^2, z + x]$ around the triangle with vertices $(0, 0, 0)$, $(1, 0, 0)$, $(1, 1, 0)$

16. $\mathbf{F} = [e^y, 0, e^x]$, C as in Prob. 15

17. $\mathbf{F} = [0, z^3, 0]$, C the boundary curve of the cylinder $x^2 + y^2 = 1, x \geqq 0, \quad y \geqq 0, \quad 0 \leqq z \leqq 1$

18. $\mathbf{F} = [-y, 2z, 0]$, C the boundary curve of $y^2 + z^2 = 4$, $z \geqq 0, \quad 0 \leqq x \leqq h$

19. $\mathbf{F} = [z, e^z, 0]$, C the boundary curve of the portion of the cone $z = \sqrt{x^2 + y^2}$, $x \geqq 0, \quad y \geqq 0, \quad 0 \leqq z \leqq 1$

20. $\mathbf{F} = [0, \cos x, 0]$, C the boundary curve of $y^2 + z^2 = 4$, $y \geqq 0, \quad z \geqq 0, \quad 0 \leqq x \leqq \pi$

CHAPTER 10 REVIEW QUESTIONS AND PROBLEMS

1. State from memory how to evaluate a line integral. A surface integral.

2. What is path independence of a line integral? What is its physical meaning and importance?

3. What line integrals can be converted to surface integrals and conversely? How?

4. What surface integrals can be converted to volume integrals? How?

5. What role did the gradient play in this chapter? The curl? State the definitions from memory.

6. What are typical applications of line integrals? Of surface integrals?

7. Where did orientation of a surface play a role? Explain.

8. State the divergence theorem from memory. Give applications.

9. In some line and surface integrals we started from vector functions, but integrands were scalar functions. How and why did we proceed in this way?

10. State Laplace's equation. Explain its physical importance. Summarize our discussion on harmonic functions.

11–20 LINE INTEGRALS (WORK INTEGRALS)

Evaluate $\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r}$ for given $\mathbf{F}$ and C by the method that seems most suitable. Remember that if $\mathbf{F}$ is a force, the integral gives the work done in the displacement along C. Show details.

11. $\mathbf{F} = [2x^2, -4y^2]$, C the straight-line segment from $(4, 2)$ to $(-6, 10)$

12. $\mathbf{F} = [y \cos xy, x \cos xy, e^z]$, C the straight-line segment from $(\pi, 1, 0)$ to $(\frac{1}{2}, \pi, 1)$

13. $\mathbf{F} = [y^2, 2xy + 5 \sin x, 0]$, C the boundary of $0 \leqq x \leqq \pi/2, \quad 0 \leqq y \leqq 2, \quad z = 0$

14. $\mathbf{F} = [-y^3, x^3 + e^{-y}, 0]$, C the circle $x^2 + y^2 = 25$, $z = 2$

15. $\mathbf{F} = [x^3, e^{2y}, e^{-4z}]$, C: $x^2 + 9y^2 = 9, \quad z = x^2$

16. $\mathbf{F} = [x^2, y^2, y^2x]$, C the helix $\mathbf{r} = [2 \cos t, 2 \sin t, 3t]$ from $(2, 0, 0)$ to $(-2, 0, 3\pi)$

17. $\mathbf{F} = [9z, 5x, 3y]$, C the ellipse $x^2 + y^2 = 9$, $z = x + 2$

18. $\mathbf{F} = [\sin \pi y, \cos \pi x, \sin \pi x]$, C the boundary curve of $0 \leqq x \leqq 1, \quad 0 \leqq y \leqq 2, \quad z = x$

19. $\mathbf{F} = [z, 2y, x]$, C the helix $r = [\cos t, \sin t, t]$ from $(1, 0, 0)$ to $(1, 0, 2\pi)$

20. $\mathbf{F} = [ze^{xz}, 2 \sinh 2y, xe^{xz}]$, C the parabola $y = x$, $z = x^2, -1 \leqq x \leqq 1$

21–25 DOUBLE INTEGRALS, CENTER OF GRAVITY

Find the coordinates $\bar{x}, \bar{y}$ of the center of gravity of a mass of density $f(x, y)$ in the region R. Sketch R, show details.

21. $f = xy$, R the triangle with vertices $(0, 0)$, $(2, 0)$, $(2, 2)$

22. $f = x^2 + y^2$, R: $x^2 + y^2 \leqq a^2$, $y \geqq 0$

23. $f = x^2$, R: $-1 \leqq x \leqq 2$, $x^2 \leqq y \leqq x + 2$. Why is $\bar{x} > 0$?

24. $f = 1$, R: $0 \leqq y \leqq 1 - x^4$

25. $f = ky$, $k > 0$, arbitrary, $0 \leqq y \leqq 1 - x^2$, $0 \leqq x \leqq 1$

26. Why are $\bar{x}$ and $\bar{y}$ in Prob. 25 independent of k?

27–35 SURFACE INTEGRALS $\iint_S \mathbf{F} \cdot \mathbf{n}\, dA$. DIVERGENCE THEOREM

Evaluate the integral diectly or, if possible, by the divergence theorem. Show details.

27. $\mathbf{F} = [ax, by, cz]$, S the sphere $x^2 + y^2 + z^2 = 36$

28. $\mathbf{F} = [x + y^2, y + z^2, z + x^2]$, S the ellipsoid with semi-axes of lengths a, b, c

29. $\mathbf{F} = [y + z, 20y, 2z^3]$, S the surface of $0 \leqq x \leqq 2$, $0 \leqq y \leqq 1, \quad 0 \leqq z \leqq y$

30. $\mathbf{F} = [1, 1, 1]$, S: $x^2 + y^2 + 4z^2 = 4, \quad z \geqq 0$

31. $\mathbf{F} = [e^x, e^y, e^z]$, S the surface of the box $|x| \leqq 1$, $|y| \leqq 1, \quad |z| \leqq 1$

32. $\mathbf{F} = [y^2, x^2, z^2]$, S the portion of the paraboloid $z = x^2 + y^2$, $z \leqq 9$

33. $\mathbf{F} = [y^2, x^2, z^2]$, S: $r = [u, u^2, v]$, $0 \leqq u \leqq 2$, $-2 \leqq v \leqq 2$

34. $\mathbf{F} = [x, xy, z]$, S the boundary of $x^2 + y^2 \leqq 1$, $0 \leqq z \leqq 5$

35. $\mathbf{F} = [x + z, y + z, x + y]$, S the sphere of radius 3 with center 0

SUMMARY OF CHAPTER 10
Vector Integral Calculus. Integral Theorems

Chapter 9 extended *differential* calculus to vectors, that is, to vector functions $\mathbf{v}(x, y, z)$ or $\mathbf{v}(t)$. Similarly, Chapter 10 extends *integral* calculus to vector functions. This involves *line integrals* (Sec. 10.1), *double integrals* (Sec. 10.3), *surface integrals* (Sec. 10.6), and *triple integrals* (Sec. 10.7) and the three "big" theorems for transforming these integrals into one another, the theorems of Green (Sec. 10.4), Gauss (Sec. 10.7), and Stokes (Sec. 10.9).

The analog of the definite integral of calculus is the **line integral** (Sec. 10.1)

$$\int_C \mathbf{F}(\mathbf{r}) \bullet d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) = \int_a^b \mathbf{F}(\mathbf{r}(t)) \bullet \frac{d\mathbf{r}}{dt}\,dt \tag{1}$$

where C: $\mathbf{r}(t) = [x(t), y(t), z(t)] = x(t)\mathbf{i} + y(t)\mathbf{j} + z(t)\mathbf{k}$ $(a \leqq t \leqq b)$ is a curve in space (or in the plane). Physically, (1) may represent the work done by a (variable) force in a displacement. Other kinds of line integrals and their applications are also discussed in Sec. 10.1.

Independence of path of a line integral in a domain D means that the integral of a given function over any path C with endpoints P and Q has the same value for all paths from P to Q that lie in D; here P and Q are fixed. An integral (1) is independent of path in D if and only if the differential form $F_1\,dx + F_2\,dy + F_3\,dz$ with continuous F_1, F_2, F_3 is **exact** in D (Sec. 10.2). Also, if $\operatorname{curl} \mathbf{F} = \mathbf{0}$, where $\mathbf{F} = [F_1, F_2, F_3]$, has continuous first partial derivatives in a *simply connected* domain D, then the integral (1) is independent of path in D (Sec. 10.2).

Integral Theorems. The formula of **Green's theorem in the plane** (Sec. 10.4)

$$\iint_R \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y}\right) dx\,dy = \oint_C (F_1\,dx + F_2\,dy) \tag{2}$$

transforms **double integrals** over a region R in the xy-plane into line integrals over the boundary curve C of R and conversely. For other forms of (2) see Sec. 10.4.

Similarly, the formula of the **divergence theorem of Gauss** (Sec. 10.7)

$$\iiint_T \operatorname{div} \mathbf{F}\,dV = \iint_S \mathbf{F} \bullet \mathbf{n}\,dA \tag{3}$$

transforms **triple integrals** over a region T in space into surface integrals over the boundary surface S of T, and conversely. Formula (3) implies **Green's formulas**

(4) $$\iiint_T (f\nabla^2 g + \nabla f \bullet \nabla g)\, dV = \iint_S f \frac{\partial g}{\partial n}\, dA,$$

(5) $$\iiint_T (f\nabla^2 g - g\nabla^2 f)\, dV = \iint_S \left(f \frac{\partial g}{\partial n} - g \frac{\partial f}{\partial n}\right) dA.$$

Finally, the formula of **Stokes's theorem** (Sec. 10.9)

(6) $$\iint_S (\text{curl } \mathbf{F}) \bullet \mathbf{n}\, dA = \oint_C \mathbf{F} \bullet \mathbf{r}'(s)\, ds$$

transforms **surface integrals** over a surface S into line integrals over the boundary curve C of S and conversely.

PART C

Fourier Analysis. Partial Differential Equations (PDEs)

CHAPTER 11 Fourier Analysis

CHAPTER 12 Partial Differential Equations (PDEs)

Chapter 11 and Chapter 12 are directly related to each other in that **Fourier analysis** has its most important applications in modeling and solving partial differential equations (PDEs) related to boundary and initial value problems of mechanics, heat flow, electrostatics, and other fields. However, the study of PDEs is a study in its own right. Indeed, PDEs are the subject of much ongoing research.

Fourier analysis allows us to model periodic phenomena which appear frequently in engineering and elsewhere—think of rotating parts of machines, alternating electric currents or the motion of planets. Related period functions may be complicated. Now, the ingenious idea of Fourier analysis is to represent complicated functions in terms of simple periodic functions, namely cosines and sines. The representations will be infinite series called **Fourier series**.[1] This idea can be generalized to more general series (see Sec. 11.5) and to integral representations (see Sec. 11.7).

The discovery of Fourier series had a huge impetus on applied mathematics as well as on mathematics as a whole. Indeed, its influence on the concept of a function, on integration theory, on convergence theory, and other theories of mathematics has been substantial (see [GenRef7] in App. 1).

Chapter 12 deals with the most important partial differential equations (PDEs) of physics and engineering, such as the wave equation, the heat equation, and the Laplace equation. These equations can model a vibrating string/membrane, temperatures on a bar, and electrostatic potentials, respectively. PDEs are very important in many areas of physics and engineering and have many more applications than ODEs.

---

[1] JEAN-BAPTISTE JOSEPH FOURIER (1768–1830), French physicist and mathematician, lived and taught in Paris, accompanied Napoléon in the Egyptian War, and was later made prefect of Grenoble. The beginnings on Fourier series can be found in works by Euler and by Daniel Bernoulli, but it was Fourier who employed them in a systematic and general manner in his main work, *Théorie analytique de la chaleur* (*Analytic Theory of Heat,* Paris, 1822), in which he developed the theory of heat conduction (heat equation; see Sec. 12.5), making these series a most important tool in applied mathematics.

Fourier Analysis

This chapter on Fourier analysis covers three broad areas: Fourier series in Secs. 11.1–11.4, more general orthonormal series called Sturm–Liouville expansions in Secs. 11.5 and 11.6 and Fourier integrals and transforms in Secs. 11.7–11.9.

The central starting point of Fourier analysis is **Fourier series**. They are infinite series designed to represent general periodic functions in terms of simple ones, namely, cosines and sines. This trigonometric system is ***orthogonal***, allowing the computation of the coefficients of the Fourier series by use of the well-known Euler formulas, as shown in Sec. 11.1. Fourier series are very important to the engineer and physicist because they allow the solution of ODEs in connection with forced oscillations (Sec. 11.3) and the approximation of periodic functions (Sec. 11.4). Moreover, applications of Fourier analysis to PDEs are given in Chap. 12. Fourier series are, in a certain sense, more universal than the familiar Taylor series in calculus because many *discontinuous* periodic functions that come up in applications can be developed in Fourier series but do not have Taylor series expansions.

The underlying idea of the Fourier series can be extended in two important ways. We can replace the trigonometric system by other families of orthogonal functions, e.g., Bessel functions and obtain the **Sturm–Liouville expansions**. Note that related Secs. 11.5 and 11.6 used to be part of Chap. 5 but, for greater readability and logical coherence, are now part of Chap. 11. The second expansion is applying Fourier series to nonperiodic phenomena and obtaining Fourier integrals and Fourier transforms. Both extensions have important applications to solving PDEs as will be shown in Chap. 12.

In a digital age, the ***discrete Fourier transform*** plays an important role. Signals, such as voice or music, are sampled and analyzed for frequencies. An important algorithm, in this context, is the ***fast Fourier transform***. This is discussed in Sec. 11.9.

Note that the two extensions of Fourier series are independent of each other and may be studied in the order suggested in this chapter or by studying Fourier integrals and transforms first and then Sturm–Liouville expansions.

Prerequisite: Elementary integral calculus (needed for Fourier coefficients).
Sections that may be omitted in a shorter course: 11.4–11.9.
References and Answers to Problems: App. 1 Part C, App. 2.

11.1 Fourier Series

Fourier series are infinite series that represent periodic functions in terms of cosines and sines. As such, Fourier series are of greatest importance to the engineer and applied mathematician. To define Fourier series, we first need some background material. A function $f(x)$ is called a **periodic function** if $f(x)$ is defined for all real x, except

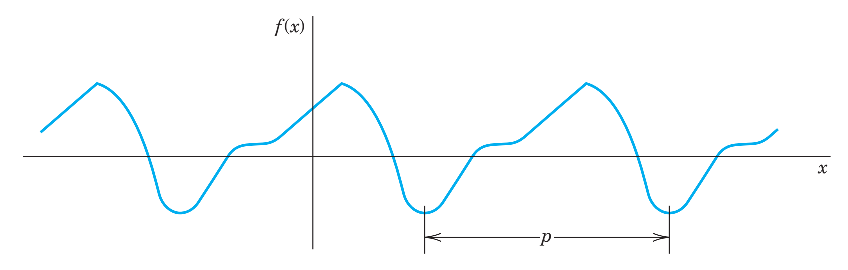


Fig. 258. Periodic function of period p

possibly at some points, and if there is some positive number p, called a **period** of $f(x)$, such that

$$f(x + p) = f(x) \qquad \text{for all } x. \tag{1}$$

(The function $f(x) = \tan x$ is a periodic function that is not defined for all real x but undefined for some points (more precisely, countably many points), that is $x = \pm\pi/2, \pm 3\pi/2, \cdots$.)

The graph of a periodic function has the characteristic that it can be obtained by periodic repetition of its graph in any interval of length p (Fig. 258).

The smallest positive period is often called the *fundamental period.* (See Probs. 2–4.)

Familiar periodic functions are the cosine, sine, tangent, and cotangent. Examples of functions that are not periodic are x, x^2, x^3, e^x, $\cosh x$, and $\ln x$, to mention just a few.

If $f(x)$ has period p, it also has the period $2p$ because (1) implies $f(x + 2p) = f([x + p] + p) = f(x + p) = f(x)$, etc.; thus for any integer $n = 1, 2, 3, \cdots$,

$$f(x + np) = f(x) \qquad \text{for all } x. \tag{2}$$

Furthermore if $f(x)$ and $g(x)$ have period p, then $af(x) + bg(x)$ with any constants a and b also has the period p.

Our problem in the first few sections of this chapter will be the representation of various ***functions $f(x)$ of period 2π*** in terms of the simple functions

$$1, \quad \cos x, \quad \sin x, \quad \cos 2x, \quad \sin 2x, \cdots, \quad \cos nx, \quad \sin nx, \cdots. \tag{3}$$

All these functions have the period 2π. They form the so-called **trigonometric system**. Figure 259 shows the first few of them (except for the constant 1, which is periodic with any period).

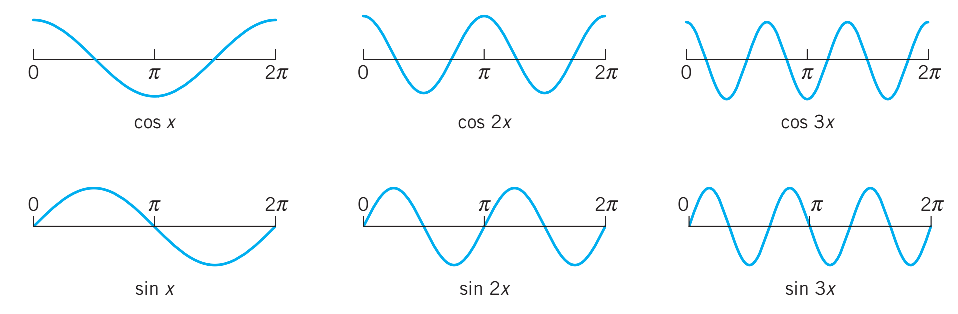


Fig. 259. Cosine and sine functions having the period 2π (the first few members of the trigonometric system (3), except for the constant 1)

The series to be obtained will be a **trigonometric series**, that is, a series of the form

$$(4) \quad \begin{aligned} &a_0 + a_1 \cos x + b_1 \sin x + a_2 \cos 2x + b_2 \sin 2x + \cdots \\ &= a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx). \end{aligned}$$

$a_0, a_1, b_1, a_2, b_2, \cdots$ are constants, called the **coefficients** of the series. We see that each term has the period 2π. Hence *if the coefficients are such that the series converges, its sum will be a function of period* 2π.

Expressions such as (4) will occur frequently in Fourier analysis. To compare the expression on the right with that on the left, simply write the terms in the summation. Convergence of one side implies convergence of the other and the sums will be the same.

Now suppose that $f(x)$ is a given function of period 2π and is such that it can be **represented** by a series (4), that is, (4) converges and, moreover, has the sum $f(x)$. Then, using the equality sign, we write

$$\textbf{(5)} \qquad f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

and call (5) the **Fourier series** of $f(x)$. We shall prove that in this case the coefficients of (5) are the so-called **Fourier coefficients** of $f(x)$, given by the **Euler formulas**

$$\textbf{(6)} \qquad \begin{aligned} &\textbf{(0)} \quad a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx \\ &\textbf{(a)} \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\, dx \qquad n = 1, 2, \cdots \\ &\textbf{(b)} \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\, dx \qquad n = 1, 2, \cdots . \end{aligned}$$

The name "Fourier series" is sometimes also used in the exceptional case that (5) with coefficients (6) does not converge or does not have the sum $f(x)$—this may happen but is merely of theoretical interest. (For Euler see footnote 4 in Sec. 2.5.)

A Basic Example

Before we derive the Euler formulas (6), let us consider how (5) and (6) are applied in this important basic example. Be fully alert, as the way we approach and solve this example will be the technique you will use for other functions. Note that the integration is a little bit different from what you are familiar with in calculus because of the n. Do not just routinely use your software but try to get a good understanding and make observations: How are continuous functions (cosines and sines) able to represent a given discontinuous function? How does the quality of the approximation increase if you take more and more terms of the series? Why are the approximating functions, called the

partial sums of the series, in this example always zero at 0 and π? Why is the factor $1/n$ (obtained in the integration) important?

EXAMPLE 1 Periodic Rectangular Wave (Fig. 260)

Find the Fourier coefficients of the periodic function $f(x)$ in Fig. 260. The formula is

$$f(x) = \begin{cases} -k & \text{if} \quad -\pi < x < 0 \\ k & \text{if} \quad 0 < x < \pi \end{cases} \qquad \text{and} \qquad f(x + 2\pi) = f(x). \tag{7}$$

Functions of this kind occur as external forces acting on mechanical systems, electromotive forces in electric circuits, etc. (The value of $f(x)$ at a single point does not affect the integral; hence we can leave $f(x)$ undefined at $x = 0$ and $x = \pm\pi$.)

Solution. From (6.0) we obtain $a_0 = 0$. This can also be seen without integration, since the area under the curve of $f(x)$ between $-\pi$ and π (taken with a minus sign where $f(x)$ is negative) is zero. From (6a) we obtain the coefficients $a_1, a_2, \cdots$ of the cosine terms. Since $f(x)$ is given by two expressions, the integrals from $-\pi$ to π split into two integrals:

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\, dx = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-k)\cos nx\, dx + \int_{0}^{\pi} k\cos nx\, dx\right]$$

$$= \frac{1}{\pi}\left[-k\frac{\sin nx}{n}\Big|_{-\pi}^{0} + k\frac{\sin nx}{n}\Big|_{0}^{\pi}\right] = 0$$

because $\sin nx = 0$ at $-\pi$, 0, and π for all $n = 1, 2, \cdots$. We see that all these cosine coefficients are zero. That is, the Fourier series of (7) has no cosine terms, just sine terms, it is a **Fourier sine series** with coefficients $b_1, b_2, \cdots$ obtained from (6b);

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin nx\, dx = \frac{1}{\pi}\left[\int_{-\pi}^{0}(-k)\sin nx\, dx + \int_{0}^{\pi} k\sin nx\, dx\right]$$

$$= \frac{1}{\pi}\left[k\frac{\cos nx}{n}\Big|_{-\pi}^{0} - k\frac{\cos nx}{n}\Big|_{0}^{\pi}\right].$$

Since $\cos(-\alpha) = \cos\alpha$ and $\cos 0 = 1$, this yields

$$b_n = \frac{k}{n\pi}[\cos 0 - \cos(-n\pi) - \cos n\pi + \cos 0] = \frac{2k}{n\pi}(1 - \cos n\pi).$$

Now, $\cos\pi = -1$, $\cos 2\pi = 1$, $\cos 3\pi = -1$, etc.; in general,

$$\cos n\pi = \begin{cases} -1 & \text{for odd } n, \\ 1 & \text{for even } n, \end{cases} \qquad \text{and thus} \qquad 1 - \cos n\pi = \begin{cases} 2 & \text{for odd } n, \\ 0 & \text{for even } n. \end{cases}$$

Hence the Fourier coefficients b_n of our function are

$$b_1 = \frac{4k}{\pi}, \qquad b_2 = 0, \qquad b_3 = \frac{4k}{3\pi}, \qquad b_4 = 0, \qquad b_5 = \frac{4k}{5\pi}, \cdots.$$

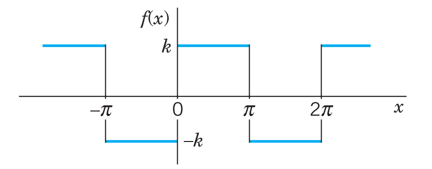


Fig. 260. Given function $f(x)$ (Periodic reactangular wave)

Since the a_n are zero, the Fourier series of $f(x)$ is

$$\frac{4k}{\pi}\left(\sin x + \tfrac{1}{3}\sin 3x + \tfrac{1}{5}\sin 5x + \cdots\right). \tag{8}$$

The partial sums are

$$S_1 = \frac{4k}{\pi}\sin x, \qquad S_2 = \frac{4k}{\pi}\left(\sin x + \frac{1}{3}\sin 3x\right). \qquad \text{etc.}$$

Their graphs in Fig. 261 seem to indicate that the series is convergent and has the sum $f(x)$, the given function. We notice that at $x = 0$ and $x = \pi$, the points of discontinuity of $f(x)$, all partial sums have the value zero, the arithmetic mean of the limits $-k$ and k of our function, at these points. This is typical.

Furthermore, assuming that $f(x)$ is the sum of the series and setting $x = \pi/2$, we have

$$f\left(\frac{\pi}{2}\right) = k = \frac{4k}{\pi}\left(1 - \frac{1}{3} + \frac{1}{5} - + \cdots\right).$$

Thus

$$1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + - \cdots = \frac{\pi}{4}.$$

This is a famous result obtained by Leibniz in 1673 from geometric considerations. It illustrates that the values of various series with constant terms can be obtained by evaluating Fourier series at specific points. ■

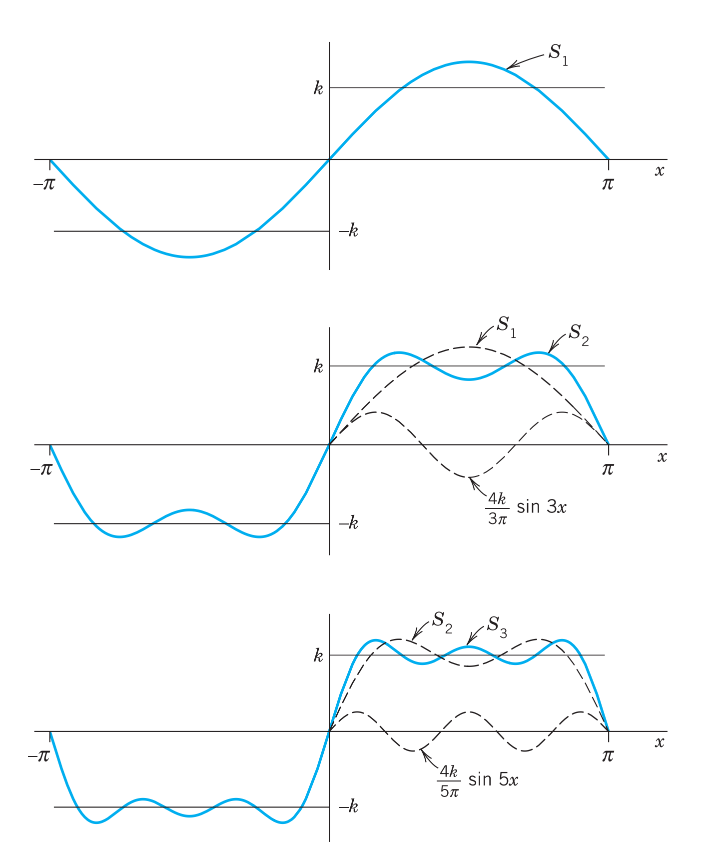


Fig. 261. First three partial sums of the corresponding Fourier series

Derivation of the Euler Formulas (6)

The key to the Euler formulas (6) is the **orthogonality** of (3), a concept of basic importance, as follows. Here we generalize the concept of inner product (Sec. 9.3) to functions.

THEOREM 1

Orthogonality of the Trigonometric System (3)

The trigonometric system (3) *is orthogonal on the interval* $-\pi \leqq x \leqq \pi$ (hence also on $0 \leqq x \leqq 2\pi$ or any other interval of length 2π because of periodicity); *that is, the integral of the product of any two functions in* (3) *over that interval is* 0, *so that for any integers n and m,*

$$\text{(9)}\quad \begin{aligned} &\textbf{(a)} \quad \int_{-\pi}^{\pi} \cos nx \cos mx \, dx = 0 \qquad (n \neq m) \\ &\textbf{(b)} \quad \int_{-\pi}^{\pi} \sin nx \sin mx \, dx = 0 \qquad (n \neq m) \\ &\textbf{(c)} \quad \int_{-\pi}^{\pi} \sin nx \cos mx \, dx = 0 \qquad (n \neq m \text{ or } n = m). \end{aligned}$$

PROOF This follows simply by transforming the integrands trigonometrically from products into sums. In (9a) and (9b), by (11) in App. A3.1,

$$\int_{-\pi}^{\pi} \cos nx \cos mx \, dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos (n+m)x \, dx + \frac{1}{2}\int_{-\pi}^{\pi} \cos (n-m)x \, dx$$

$$\int_{-\pi}^{\pi} \sin nx \sin mx \, dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos (n-m)x \, dx - \frac{1}{2}\int_{-\pi}^{\pi} \cos (n+m)x \, dx.$$

Since $m \neq n$ (integer!), the integrals on the right are all 0. Similarly, in (9c), for all integer m and n (without exception; do you see why?)

$$\int_{-\pi}^{\pi} \sin nx \cos mx \, dx = \frac{1}{2}\int_{-\pi}^{\pi} \sin (n+m)x \, dx + \frac{1}{2}\int_{-\pi}^{\pi} \sin (n-m)x \, dx = 0 + 0. \qquad \blacksquare$$

Application of Theorem 1 to the Fourier Series (5)

We prove (6.0). Integrating on both sides of (5) from $-\pi$ to π, we get

$$\int_{-\pi}^{\pi} f(x) \, dx = \int_{-\pi}^{\pi} \left[a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx) \right] dx.$$

We now assume that termwise integration is allowed. (We shall say in the proof of Theorem 2 when this is true.) Then we obtain

$$\int_{-\pi}^{\pi} f(x) \, dx = a_0 \int_{-\pi}^{\pi} dx + \sum_{n=1}^{\infty} \left(a_n \int_{-\pi}^{\pi} \cos nx \, dx + b_n \int_{-\pi}^{\pi} \sin nx \, dx \right).$$

The first term on the right equals $2\pi a_0$. Integration shows that all the other integrals are 0. Hence division by 2π gives (6.0).

We prove (6a). Multiplying (5) on both sides by $\cos mx$ with any ***fixed*** positive integer m and integrating from $-\pi$ to π, we have

$$(10)\qquad \int_{-\pi}^{\pi} f(x)\cos mx\,dx = \int_{-\pi}^{\pi}\left[a_0 + \sum_{n=1}^{\infty}(a_n\cos nx + b_n\sin nx)\right]\cos mx\,dx.$$

We now integrate term by term. Then on the right we obtain an integral of $a_0 \cos mx$, which is 0; an integral of $a_n \cos nx \cos mx$, which is $a_m\pi$ for $n = m$ and 0 for $n \neq m$ by (9a); and an integral of $b_n \sin nx \cos mx$, which is 0 for all n and m by (9c). Hence the right side of (10) equals $a_m\pi$. Division by π gives (6a) (with m instead of n).

We finally prove (6b). Multiplying (5) on both sides by $\sin mx$ with any ***fixed*** positive integer m and integrating from $-\pi$ to π, we get

$$(11)\qquad \int_{-\pi}^{\pi} f(x)\sin mx\,dx = \int_{-\pi}^{\pi}\left[a_0 + \sum_{n=1}^{\infty}(a_n\cos nx + b_n\sin nx)\right]\sin mx\,dx.$$

Integrating term by term, we obtain on the right an integral of $a_0 \sin mx$, which is 0; an integral of $a_n \cos nx \sin mx$, which is 0 by (9c); and an integral of $b_n \sin nx \sin mx$, which is $b_m\pi$ if $n = m$ and 0 if $n \neq m$, by (9b). This implies (6b) (with n denoted by m). This completes the proof of the Euler formulas (6) for the Fourier coefficients. ■

Convergence and Sum of a Fourier Series

The class of functions that can be represented by Fourier series is surprisingly large and general. Sufficient conditions valid in most applications are as follows.

THEOREM 2

Representation by a Fourier Series

Let $f(x)$ be periodic with period 2π and piecewise continuous (see Sec. 6.1) *in the interval $-\pi \leqq x \leqq \pi$. Furthermore, let $f(x)$ have a left-hand derivative and a right-hand derivative at each point of that interval. Then the Fourier series* (5) *of $f(x)$* [with coefficients (6)] *converges. Its sum is $f(x)$, except at points x_0 where $f(x)$ is discontinuous. There the sum of the series is the average of the left- and right-hand limits*[2] *of $f(x)$ at x_0.*

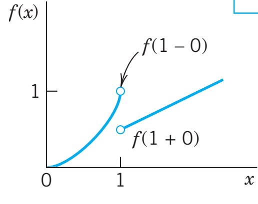


Fig. 262. Left- and right-hand limits

$f(1-0) = 1,$

$f(1+0) = \frac{1}{2}$

of the function

$$f(x) = \begin{cases} x^2 & \text{if } x < 1 \\ x/2 & \text{if } x \geqq 1 \end{cases}$$

[2]The **left-hand limit** of $f(x)$ at x_0 is defined as the limit of $f(x)$ as x approaches x_0 from the left and is commonly denoted by $f(x_0 - 0)$. Thus

$$f(x_0 - 0) = \lim_{h\to 0} f(x_0 - h) \text{ as } h \to 0 \text{ through positive values.}$$

The **right-hand limit** is denoted by $f(x_0 + 0)$ and

$$f(x_0 + 0) = \lim_{h\to 0} f(x_0 + h) \text{ as } h \to 0 \text{ through positive values.}$$

The **left-** and **right-hand derivatives** of $f(x)$ at x_0 are defined as the limits of

$$\frac{f(x_0 - h) - f(x_0 - 0)}{-h} \quad \text{and} \quad \frac{f(x_0 + h) - f(x_0 + 0)}{-h},$$

respectively, as $h \to 0$ through positive values. Of course if $f(x)$ is continuous at x_0, the last term in both numerators is simply $f(x_0)$.

PROOF We prove convergence, but only for a continuous function $f(x)$ having continuous first and second derivatives. And we do not prove that the sum of the series is $f(x)$ because these proofs are much more advanced; see, for instance, Ref. [C12] listed in App. 1. Integrating (6a) by parts, we obtain

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos nx\,dx = \frac{f(x)\sin nx}{n\pi}\Bigg|_{-\pi}^{\pi} - \frac{1}{n\pi}\int_{-\pi}^{\pi} f'(x)\sin nx\,dx.$$

The first term on the right is zero. Another integration by parts gives

$$a_n = \frac{f'(x)\cos nx}{n^2\pi}\Bigg|_{-\pi}^{\pi} - \frac{1}{n^2\pi}\int_{-\pi}^{\pi} f''(x)\cos nx\,dx.$$

The first term on the right is zero because of the periodicity and continuity of $f'(x)$. Since f'' is continuous in the interval of integration, we have

$$|f''(x)| < M$$

for an appropriate constant M. Furthermore, $|\cos nx| \leqq 1$. It follows that

$$|a_n| = \frac{1}{n^2\pi}\left|\int_{-\pi}^{\pi} f''(x)\cos nx\,dx\right| < \frac{1}{n^2\pi}\int_{-\pi}^{\pi} M\,dx = \frac{2M}{n^2}.$$

Similarly, $|b_n| < 2\,M/n^2$ for all n. Hence the absolute value of each term of the Fourier series of $f(x)$ is at most equal to the corresponding term of the series

$$|a_0| + 2M\left(1 + 1 + \frac{1}{2^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{3^2} + \cdots\right)$$

which is convergent. Hence that Fourier series converges and the proof is complete. (Readers already familiar with uniform convergence will see that, by the Weierstrass test in Sec. 15.5, under our present assumptions the Fourier series converges uniformly, and our derivation of (6) by integrating term by term is then justified by Theorem 3 of Sec. 15.5.) ■

EXAMPLE 2 **Convergence at a Jump as Indicated in Theorem 2**

The rectangular wave in Example 1 has a jump at $x = 0$. Its left-hand limit there is $-k$ and its right-hand limit is k (Fig. 261). Hence the average of these limits is 0. The Fourier series (8) of the wave does indeed converge to this value when $x = 0$ because then all its terms are 0. Similarly for the other jumps. This is in agreement with Theorem 2. ■

Summary. A Fourier series of a given function $f(x)$ of period 2π is a series of the form (5) with coefficients given by the Euler formulas (6). Theorem 2 gives conditions that are sufficient for this series to converge and at each x to have the value $f(x)$, except at discontinuities of $f(x)$, where the series equals the arithmetic mean of the left-hand and right-hand limits of $f(x)$ at that point.

PROBLEM SET 11.1

1–5 PERIOD, FUNDAMENTAL PERIOD

The *fundamental period* is the smallest positive period. Find it for

1. $\cos x$, $\sin x$, $\cos 2x$, $\sin 2x$, $\cos \pi x$, $\sin \pi x$, $\cos 2\pi x$, $\sin 2\pi x$
2. $\cos nx$, $\sin nx$, $\cos \frac{2\pi x}{k}$, $\sin \frac{2\pi x}{k}$, $\cos \frac{2\pi nx}{k}$, $\sin \frac{2\pi nx}{k}$
3. If $f(x)$ and $g(x)$ have period p, show that $h(x) = af(x) + bg(x)$ (a, b, constant) has the period p. Thus all functions of period p form a **vector space**.
4. **Change of scale.** If $f(x)$ has period p, show that $f(ax)$, $a \neq 0$, and $f(x/b)$, $b \neq 0$, are periodic functions of x of periods p/a and bp, respectively. Give examples.
5. Show that $f =$ const is periodic with any period but has no fundamental period.

6–10 GRAPHS OF 2π–PERIODIC FUNCTIONS

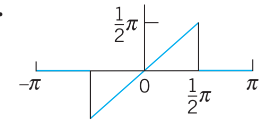
Sketch or graph $f(x)$ which for $-\pi < x < \pi$ is given as follows.

6. $f(x) = |x|$
7. $f(x) = |\sin x|$, $f(x) = \sin |x|$
8. $f(x) = e^{-|x|}$, $f(x) = |e^{-x}|$
9. $f(x) = \begin{cases} x & \text{if } -\pi < x < 0 \\ \pi - x & \text{if } 0 < x < \pi \end{cases}$
10. $f(x) = \begin{cases} -\cos^2 x & \text{if } -\pi < x < 0 \\ \cos^2 x & \text{if } 0 < x < \pi \end{cases}$
11. **Calculus review.** Review integration techniques for integrals as they are likely to arise from the Euler formulas, for instance, definite integrals of $x \cos nx$, $x^2 \sin nx$, $e^{-2x} \cos nx$, etc.

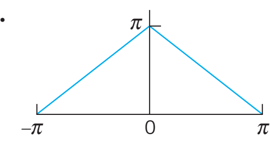
12–21 FOURIER SERIES

Find the Fourier series of the given function $f(x)$, which is assumed to have the period 2π. Show the details of your work. Sketch or graph the partial sums up to that including $\cos 5x$ and $\sin 5x$.

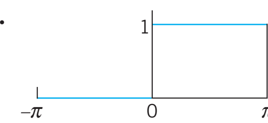
12. $f(x)$ in Prob. 6
13. $f(x)$ in Prob. 9
14. $f(x) = x^2$ $(-\pi < x < \pi)$
15. $f(x) = x^2$ $(0 < x < 2\pi)$
16.



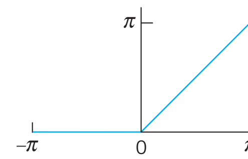
17.



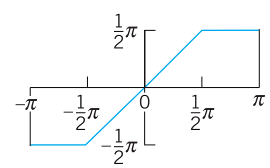
18.



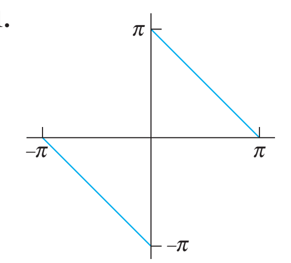
19.



20.



21.



22. **CAS EXPERIMENT. Graphing.** Write a program for graphing partial sums of the following series. Guess from the graph what $f(x)$ the series may represent. Confirm or disprove your guess by using the Euler formulas.
    (a) $2(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots)$
    $\quad - 2(\frac{1}{2}\sin 2x + \frac{1}{4}\sin 4x + \frac{1}{6}\sin 6x \cdots)$
    (b) $\frac{1}{2} + \frac{4}{\pi^2}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$
    (c) $\frac{2}{3}\pi^2 + 4(\cos x - \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x - \frac{1}{16}\cos 4x + - \cdots)$
23. **Discontinuities.** Verify the last statement in Theorem 2 for the discontinuities of $f(x)$ in Prob. 21.
24. **CAS EXPERIMENT. Orthogonality.** Integrate and graph the integral of the product $\cos mx \cos nx$ (with various integer m and n of your choice) from $-a$ to a as a function of a and conclude orthogonality of $\cos mx$

and cos nx ($m \neq n$) for $a = \pi$ from the graph. For what m and n will you get orthogonality for $a = \pi/2, \pi/3, \pi/4$? Other a? Extend the experiment to cos mx sin nx and sin mx sin nx.

25. CAS EXPERIMENT. Order of Fourier Coefficients. The order seems to be $1/n$ if f is discontinous, and $1/n^2$ if f is continuous but $f' = df/dx$ is discontinuous, $1/n^3$ if f and f' are continuous but f'' is discontinuous, etc. Try to verify this for examples. Try to prove it by integrating the Euler formulas by parts. What is the practical significance of this?

11.2 Arbitrary Period. Even and Odd Functions. Half-Range Expansions

We now expand our initial basic discussion of Fourier series.

Orientation. This section concerns three topics:

1. Transition from period 2π to any period $2L$, for the function f, simply by a transformation of scale on the x-axis.
2. Simplifications. Only cosine terms if f is even ("Fourier cosine series"). Only sine terms if f is odd ("Fourier sine series").
3. Expansion of f given for $0 \leqq x \leqq L$ in two Fourier series, one having only cosine terms and the other only sine terms ("half-range expansions").

1. From Period 2π to Any Period $p = 2L$

Clearly, periodic functions in applications may have any period, not just 2π as in the last section (chosen to have simple formulas). The notation $p = 2L$ for the period is practical because L will be a length of a violin string in Sec. 12.2, of a rod in heat conduction in Sec. 12.5, and so on.

The transition from period 2π to be period $p = 2L$ is effected by a suitable change of scale, as follows. Let $f(x)$ have period $p = 2L$. Then we can introduce a new variable v such that $f(x)$, as a function of v, has period 2π. If we set

$$\text{(1)} \qquad \text{(a)} \quad x = \frac{p}{2\pi} v, \qquad \text{so that} \qquad \text{(b)} \quad v = \frac{2\pi}{p} x = \frac{\pi}{L} x$$

then $v = \pm\pi$ corresponds to $x = \pm L$. This means that f, as a function of v, has period 2π and, therefore, a Fourier series of the form

$$\text{(2)} \qquad f(x) = f\left(\frac{L}{\pi} v\right) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nv + b_n \sin nv)$$

with coefficients obtained from (6) in the last section

$$\text{(3)} \qquad \begin{aligned} a_0 &= \frac{1}{2\pi} \int_{-\pi}^{\pi} f\left(\frac{L}{\pi} v\right) dv, \qquad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f\left(\frac{L}{\pi} v\right) \cos nv \, dv, \\ b_n &= \frac{1}{\pi} \int_{-\pi}^{\pi} f\left(\frac{L}{\pi} v\right) \sin nv \, dv. \end{aligned}$$

We could use these formulas directly, but the change to x simplifies calculations. Since

$$v = \frac{\pi}{L}x, \qquad \text{we have} \qquad dv = \frac{\pi}{L}\,dx \tag{4}$$

and we integrate over x from $-L$ to L. Consequently, we obtain for a function $f(x)$ of period $2L$ the Fourier series

$$f(x) = a_0 + \sum_{n=1}^{\infty}\left(a_n \cos\frac{n\pi}{L}x + b_n \sin\frac{n\pi}{L}x\right) \tag{5}$$

with the **Fourier coefficients** of $f(x)$ given by the **Euler formulas** (π/L in dx cancels $1/\pi$ in (3))

(6)

$$\text{(0)} \quad a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\,dx$$

$$\text{(a)} \quad a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\frac{n\pi x}{L}\,dx \qquad n = 1, 2, \cdots$$

$$\text{(b)} \quad b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\frac{n\pi x}{L}\,dx \qquad n = 1, 2, \cdots.$$

Just as in Sec. 11.1, we continue to call (5) with any coefficients a **trigonometric series**. And we can integrate from 0 to $2L$ or over any other interval of length $p = 2L$.

EXAMPLE 1 Periodic Rectangular Wave

Find the Fourier series of the function (Fig. 263)

$$f(x) = \begin{cases} 0 & \text{if} \quad -2 < x < -1 \\ k & \text{if} \quad -1 < x < 1 \\ 0 & \text{if} \quad 1 < x < 2 \end{cases} \qquad p = 2L = 4, \quad L = 2.$$

Solution. From (6.0) we obtain $a_0 = k/2$ (verify!). From (6a) we obtain

$$a_n = \frac{1}{2}\int_{-2}^{2} f(x)\cos\frac{n\pi x}{2}\,dx = \frac{1}{2}\int_{-1}^{1} k\cos\frac{n\pi x}{2}\,dx = \frac{2k}{n\pi}\sin\frac{n\pi}{2}.$$

Thus $a_n = 0$ if n is even and

$$a_n = 2k/n\pi \quad \text{if} \quad n = 1, 5, 9, \cdots, \qquad a_n = -2k/n\pi \quad \text{if} \quad n = 3, 7, 11, \cdots.$$

From (6b) we find that $b_n = 0$ for $n = 1, 2, \cdots$. Hence the Fourier series is a **Fourier cosine series** (that is, it has no sine terms)

$$f(x) = \frac{k}{2} + \frac{2k}{\pi}\left(\cos\frac{\pi}{2}x - \frac{1}{3}\cos\frac{3\pi}{2}x + \frac{1}{5}\cos\frac{5\pi}{2}x - + \cdots\right).$$ ■

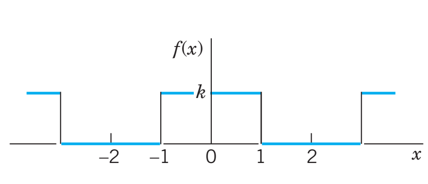


Fig. 263. Example 1

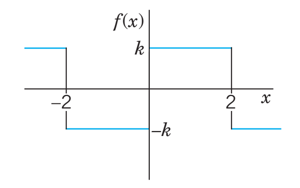


Fig. 264. Example 2

EXAMPLE 2 Periodic Rectangular Wave. Change of Scale

Find the Fourier series of the function (Fig. 264)

$$f(x) = \begin{cases} -k & \text{if} \quad -2 < x < 0 \\ k & \text{if} \quad 0 < x < 2 \end{cases} \qquad p = 2L = 4, \qquad L = 2.$$

Solution. Since $L = 2$, we have in (3) $v = \pi x/2$ and obtain from (8) in Sec. 11.1 with v instead of x, that is,

$$g(v) = \frac{4k}{\pi}\left(\sin v + \frac{1}{3}\sin 3v + \frac{1}{5}\sin 5v + \cdots\right)$$

the present Fourier series

$$f(x) = \frac{4k}{\pi}\left(\sin\frac{\pi}{2}x + \frac{1}{3}\sin\frac{3\pi}{2}x + \frac{1}{5}\sin\frac{5\pi}{2}x + \cdots\right).$$

Confirm this by using (6) and integrating. ■

EXAMPLE 3 Half-Wave Rectifier

A sinusoidal voltage $E \sin \omega t$, where t is time, is passed through a half-wave rectifier that clips the negative portion of the wave (Fig. 265). Find the Fourier series of the resulting periodic function

$$u(t) = \begin{cases} 0 & \text{if} \quad -L < t < 0, \\ E\sin\omega t & \text{if} \quad 0 < t < L \end{cases} \qquad p = 2L = \frac{2\pi}{\omega}, \qquad L = \frac{\pi}{\omega}.$$

Solution. Since $u = 0$ when $-L < t < 0$, we obtain from (6.0), with t instead of x,

$$a_0 = \frac{\omega}{2\pi}\int_0^{\pi/\omega} E\sin\omega t\, dt = \frac{E}{\pi}$$

and from (6a), by using formula (11) in App. A3.1 with $x = \omega t$ and $y = n\omega t$,

$$a_n = \frac{\omega}{\pi}\int_0^{\pi/\omega} E\sin\omega t\cos n\omega t\, dt = \frac{\omega E}{2\pi}\int_0^{\pi/\omega} [\sin(1+n)\,\omega t + \sin(1-n)\,\omega t]\, dt.$$

If $n = 1$, the integral on the right is zero, and if $n = 2, 3, \cdots$, we readily obtain

$$a_n = \frac{\omega E}{2\pi}\left[-\frac{\cos(1+n)\omega t}{(1+n)\omega} - \frac{\cos(1-n)\omega t}{(1-n)\omega}\right]_0^{\pi/\omega}$$

$$= \frac{E}{2\pi}\left(\frac{-\cos(1+n)\pi + 1}{1+n} + \frac{-\cos(1-n)\pi + 1}{1-n}\right).$$

If n is odd, this is equal to zero, and for even n we have

$$a_n = \frac{E}{2\pi}\left(\frac{2}{1+n} + \frac{2}{1-n}\right) = -\frac{2E}{(n-1)(n+1)\pi} \qquad (n = 2, 4, \cdots).$$

In a similar fashion we find from (6b) that $b_1 = E/2$ and $b_n = 0$ for $n = 2, 3, \cdots$. Consequently,

$$u(t) = \frac{E}{\pi} + \frac{E}{2}\sin\omega t - \frac{2E}{\pi}\left(\frac{1}{1\cdot 3}\cos 2\omega t + \frac{1}{3\cdot 5}\cos 4\omega t + \cdots\right).$$

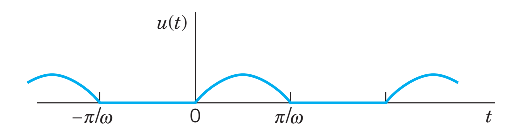


Fig. 265. Half-wave rectifier

2. Simplifications: Even and Odd Functions

Fig. 266. Even function

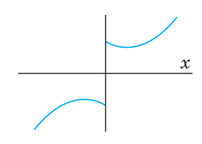


Fig. 267. Odd function

If $f(x)$ is an **even function**, that is, $f(-x) = f(x)$ (see Fig. 266), its Fourier series (5) reduces to a **Fourier cosine series**

(5*)
$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos\frac{n\pi}{L}x \qquad (f \text{ even})$$

with coefficients (note: integration from 0 to L only!)

(6*)
$$a_0 = \frac{1}{L}\int_0^L f(x)\,dx, \qquad a_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx, \qquad n = 1, 2, \cdots.$$

If $f(x)$ is an **odd function**, that is, $f(-x) = -f(x)$ (see Fig. 267), its Fourier series (5) reduces to a **Fourier sine series**

(5)**
$$f(x) = \sum_{n=1}^{\infty} b_n \sin\frac{n\pi}{L}x \qquad (f \text{ odd})$$

with coefficients

(6**)
$$b_n = \frac{2}{L}\int_0^L f(x)\sin\frac{n\pi x}{L}\,dx.$$

These formulas follow from (5) and (6) by remembering from calculus that the definite integral gives the net area (= area above the axis minus area below the axis) under the curve of a function between the limits of integration. This implies

(7)
$$\text{(a)} \quad \int_{-L}^{L} g(x)\,dx = 2\int_0^L g(x)\,dx \qquad \text{for even } g$$
$$\text{(b)} \quad \int_{-L}^{L} h(x)\,dx = 0 \qquad \text{for odd } h$$

Formula (7b) implies the reduction to the cosine series (even f makes $f(x)\sin(n\pi x/L)$ odd since sin is odd) and to the sine series (odd f makes $f(x)\cos(n\pi x/L)$ odd since cos is even). Similarly, (7a) reduces the integrals in (6*) and (6**) to integrals from 0 to L. These reductions are obvious from the graphs of an even and an odd function. (Give a formal proof.)

Summary

Even Function of Period 2π. If f is even and $L = \pi$, then

$$f(x) = a_0 + \sum_{n=1}^{\infty} a_n \cos nx$$

with coefficients

$$a_0 = \frac{1}{\pi}\int_0^{\pi} f(x)\,dx, \qquad a_n = \frac{2}{\pi}\int_0^{\pi} f(x)\cos nx\,dx, \quad n = 1, 2, \cdots$$

Odd Function of Period 2π. If f is odd and $L = \pi$, then

$$f(x) = \sum_{n=1}^{\infty} b_n \sin nx$$

with coefficients

$$b_n = \frac{2}{\pi}\int_0^{\pi} f(x)\sin nx\,dx, \qquad n = 1, 2, \cdots.$$

EXAMPLE 4 Fourier Cosine and Sine Series

The rectangular wave in Example 1 is even. Hence it follows without calculation that its Fourier series is a Fourier cosine series, the b_n are all zero. Similarly, it follows that the Fourier series of the odd function in Example 2 is a Fourier sine series.

In Example 3 you can see that the Fourier cosine series represents $u(t) - E/\pi - \frac{1}{2}E \sin \omega t$. Can you prove that this is an even function? ■

Further simplifications result from the following property, whose very simple proof is left to the student.

THEOREM 1 Sum and Scalar Multiple

The Fourier coefficients of a sum $f_1 + f_2$ are the sums of the corresponding Fourier coefficients of f_1 and f_2.

The Fourier coefficients of cf are c times the corresponding Fourier coefficients of f.

EXAMPLE 5 Sawtooth Wave

Find the Fourier series of the function (Fig. 268)

$$f(x) = x + \pi \quad \text{if} \quad -\pi < x < \pi \qquad \text{and} \qquad f(x + 2\pi) = f(x).$$

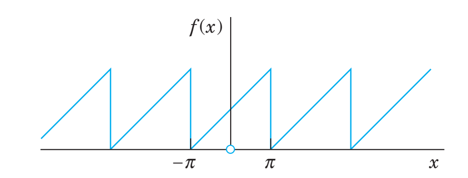


Fig. 268. The function $f(x)$. Sawtooth wave

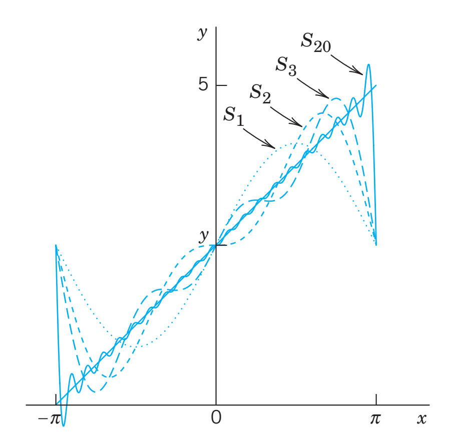


Fig. 269. Partial sums S_1, S_2, S_3, S_{20} in Example 5

Solution. We have $f = f_1 + f_2$, where $f_1 = x$ and $f_2 = \pi$. The Fourier coefficients of f_2 are zero, except for the first one (the constant term), which is π. Hence, by Theorem 1, the Fourier coefficients a_n, b_n are those of f_1, except for a_0, which is π. Since f_1 is odd, $a_n = 0$ for $n = 1, 2, \cdots$, and

$$b_n = \frac{2}{\pi}\int_0^{\pi} f_1(x)\sin nx\,dx = \frac{2}{\pi}\int_0^{\pi} x\sin nx\,dx.$$

Integrating by parts, we obtain

$$b_n = \frac{2}{\pi}\left[\frac{-x\cos nx}{n}\Bigg|_0^{\pi} + \frac{1}{n}\int_0^{\pi}\cos nx\,dx\right] = -\frac{2}{n}\cos n\pi.$$

Hence $b_1 = 2$, $b_2 = -\frac{2}{2}$, $b_3 = \frac{2}{3}$, $b_4 = -\frac{2}{4}, \cdots$, and the Fourier series of $f(x)$ is

$$f(x) = \pi + 2\left(\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots\right). \qquad \text{(Fig. 269)}$$ ■

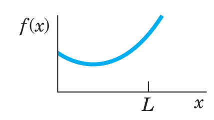
3. Half-Range Expansions

Half-range expansions are Fourier series. The idea is simple and useful. Figure 270 explains it. We want to represent $f(x)$ in Fig. 270.0 by a Fourier series, where $f(x)$ may be the shape of a distorted violin string or the temperature in a metal bar of length L, for example. (Corresponding problems will be discussed in Chap. 12.) Now comes the idea.

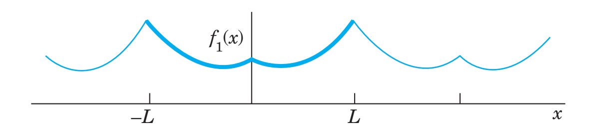
We could extend $f(x)$ as a function of period L and develop the extended function into a Fourier series. But this series would, in general, contain *both* cosine *and* sine terms. We can do better and get simpler series. Indeed, for our given f we can calculate Fourier coefficients from (6*) or from (6**). And we have a choice and can take what seems more practical. If we use (6*), we get (5*). This is the **even periodic extension** f_1 of f in Fig. 270a. If we choose (6**) instead, we get (5**), the **odd periodic extension** f_2 of f in Fig. 270b.

Both extensions have period $2L$. This motivates the name **half-range expansions**: f is given (and of physical interest) only on half the range, that is, on half the interval of periodicity of length $2L$.

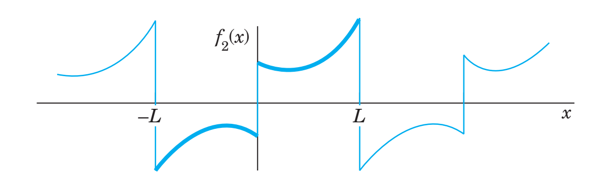
Let us illustrate these ideas with an example that we shall also need in Chap. 12.



(0) The given function $f(x)$



(a) $f(x)$ continued as an ***even*** periodic function of period $2L$



(b) $f(x)$ continued as an ***odd*** periodic function of period $2L$

Fig. 270. Even and odd extensions of period $2L$

EXAMPLE 6 "Triangle" and Its Half-Range Expansions

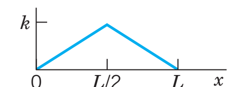


Fig. 271. The given function in Example 6

Find the two half-range expansions of the function (Fig. 271)

$$f(x) = \begin{cases} \dfrac{2k}{L}x & \text{if} \quad 0 < x < \dfrac{L}{2} \\[2ex] \dfrac{2k}{L}(L - x) & \text{if} \quad \dfrac{L}{2} < x < L. \end{cases}$$

Solution. **(a)** ***Even periodic extension.*** From (6*) we obtain

$$a_0 = \frac{1}{L}\left[\frac{2k}{L}\int_0^{L/2} x\,dx + \frac{2k}{L}\int_{L/2}^{L}(L - x)\,dx\right] = \frac{k}{2},$$

$$a_n = \frac{2}{L}\left[\frac{2k}{L}\int_L^{L/2} x\cos\frac{n\pi}{L}x\,dx + \frac{2k}{L}\int_{L/2}^{L}(L - x)\cos\frac{n\pi}{L}x\,dx\right].$$

We consider a_n. For the first integral we obtain by integration by parts

$$\int_0^{L/2} x\cos\frac{n\pi}{L}x\,dx = \frac{Lx}{n\pi}\sin\frac{n\pi}{L}x\,\Big|_0^{L/2} - \frac{L}{n\pi}\int_0^{L/2}\sin\frac{n\pi}{L}x\,dx$$

$$= \frac{L^2}{2n\pi}\sin\frac{n\pi}{2} + \frac{L^2}{n^2\pi^2}\left(\cos\frac{n\pi}{2} - 1\right).$$

Similarly, for the second integral we obtain

$$\int_{L/2}^{L}(L - x)\cos\frac{n\pi}{L}x\,dx = \frac{L}{n\pi}(L - x)\sin\frac{n\pi}{L}x\,\Big|_{L/2}^{L} + \frac{L}{n\pi}\int_{L/2}^{L}\sin\frac{n\pi}{L}x\,dx$$

$$= \left(0 - \frac{L}{n\pi}\left(L - \frac{L}{2}\right)\sin\frac{n\pi}{2}\right) - \frac{L^2}{n^2\pi^2}\left(\cos n\pi - \cos\frac{n\pi}{2}\right).$$

We insert these two results into the formula for a_n. The sine terms cancel and so does a factor L^2. This gives

$$a_n = \frac{4k}{n^2\pi^2}\left(2\cos\frac{n\pi}{2} - \cos n\pi - 1\right).$$

Thus,

$$a_2 = -16k/(2^2\pi^2), \qquad a_6 = -16k/(6^2\pi^2), \qquad a_{10} = -16k/(10^2\pi^2), \cdots$$

and $a_n = 0$ if $n \neq 2, 6, 10, 14, \cdots$. Hence the first half-range expansion of $f(x)$ is (Fig. 272a)

$$f(x) = \frac{k}{2} - \frac{16k}{\pi^2}\left(\frac{1}{2^2}\cos\frac{2\pi}{L}x + \frac{1}{6^2}\cos\frac{6\pi}{L}x + \cdots\right).$$

This Fourier cosine series represents the even periodic extension of the given function $f(x)$, of period $2L$.

(b) ***Odd periodic extension.*** Similarly, from (6**) we obtain

$$b_n = \frac{8k}{n^2\pi^2}\sin\frac{n\pi}{2}. \tag{5}$$

Hence the other half-range expansion of $f(x)$ is (Fig. 272b)

$$f(x) = \frac{8k}{\pi^2}\left(\frac{1}{1^2}\sin\frac{\pi}{L}x - \frac{1}{3^2}\sin\frac{3\pi}{L}x + \frac{1}{5^2}\sin\frac{5\pi}{L}x - + \cdots\right).$$

The series represents the odd periodic extension of $f(x)$, of period $2L$.

Basic applications of these results will be shown in Secs. 12.3 and 12.5. ■

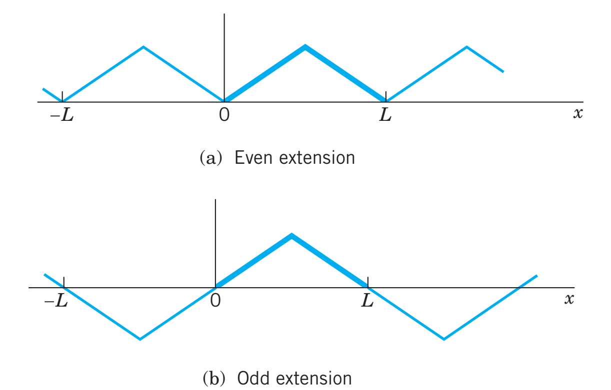


Fig. 272. Periodic extensions of $f(x)$ in Example 6

PROBLEM SET 11.2

1–7 EVEN AND ODD FUNCTIONS

Are the following functions even or odd or neither even nor odd?

1. e^x, $e^{-|x|}$, $x^3\cos nx$, $x^2\tan \pi x$, $\sinh x - \cosh x$

2. $\sin^2 x$, $\sin(x^2)$, $\ln x$, $x/(x^2+1)$, $x\cot x$

3. Sums and products of even functions

4. Sums and products of odd functions

5. Absolute values of odd functions

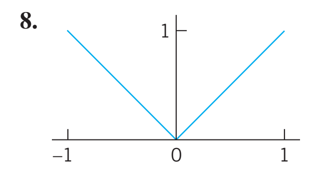
6. Product of an odd times an even function

7. Find all functions that are both even and odd.

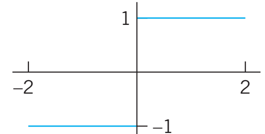
8–17 FOURIER SERIES FOR PERIOD $p = 2L$

Is the given function even or odd or neither even nor odd? Find its Fourier series. Show details of your work.

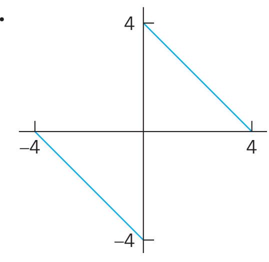
8.



9.



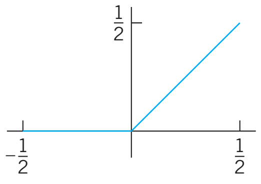
10.



11. $f(x) = x^2 \quad (-1 < x < 1), \quad p = 2$

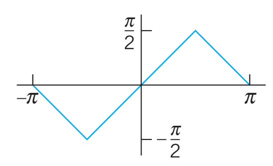
12. $f(x) = 1 - x^2/4 \quad (-2 < x < 2), \quad p = 4$

13.



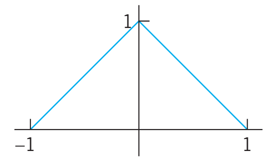
14. $f(x) = \cos \pi x \quad (-\frac{1}{2} < x < \frac{1}{2}), \quad p = 1$

15.



16. $f(x) = x|x| \quad (-1 < x < 1), \quad p = 2$

17.



18. Rectifier. Find the Fourier series of the function obtained by passing the voltage $v(t) = V_0 \cos 100\pi t$ through a half-wave rectifier that clips the negative half-waves.

19. Trigonometric Identities. Show that the familiar identities $\cos^3 x = \frac{3}{4}\cos x + \frac{1}{4}\cos 3x$ and $\sin^3 x = \frac{3}{4}\sin x - \frac{1}{4}\sin 3x$ can be interpreted as Fourier series expansions. Develop $\cos^4 x$.

20. Numeric Values. Using Prob. 11, show that $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots = \frac{1}{6}\pi^2$.

21. CAS PROJECT. Fourier Series of 2*L*-Periodic Functions. (a) Write a program for obtaining partial sums of a Fourier series (5).

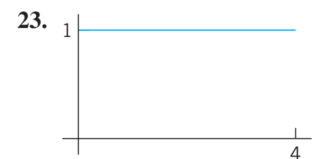
(b) Apply the program to Probs. 8–11, graphing the first few partial sums of each of the four series on common axes. Choose the first five or more partial sums until they approximate the given function reasonably well. Compare and comment.

22. Obtain the Fourier series in Prob. 8 from that in Prob. 17.

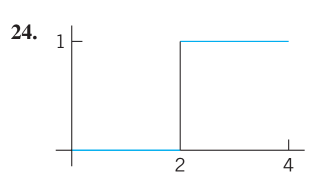
23–29 HALF-RANGE EXPANSIONS

Find **(a)** the Fourier cosine series, **(b)** the Fourier sine series. Sketch $f(x)$ and its two periodic extensions. Show the details.

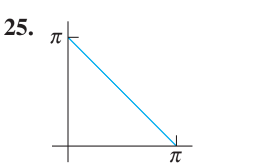
23.



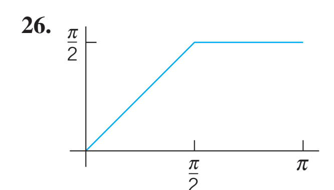
24.



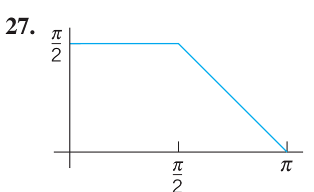
25.



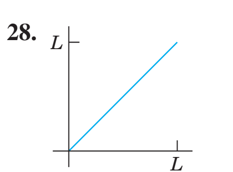
26.



27.



28.



29. $f(x) = \sin x \; (0 < x < \pi)$

30. Obtain the solution to Prob. 26 from that of Prob. 27.

11.3 Forced Oscillations

Fourier series have important applications for both ODEs and PDEs. In this section we shall focus on ODEs and cover similar applications for PDEs in Chap. 12. All these applications will show our indebtedness to Euler's and Fourier's ingenious idea of splitting up periodic functions into the simplest ones possible.

From Sec. 2.8 we know that forced oscillations of a body of mass m on a spring of modulus k are governed by the ODE

(1) $$my'' + cy' + ky = r(t)$$

where $y = y(t)$ is the displacement from rest, c the damping constant, k the spring constant (spring modulus), and $r(t)$ the external force depending on time t. Figure 274 shows the model and Fig. 275 its electrical analog, an *RLC*-circuit governed by

(1*) $$LI'' + RI' + \frac{1}{C}I = E'(t) \qquad \text{(Sec. 2.9).}$$

We consider (1). If $r(t)$ is a sine or cosine function and if there is damping ($c > 0$), then the steady-state solution is a harmonic oscillation with frequency equal to that of $r(t)$. However, if $r(t)$ is not a pure sine or cosine function but is any other periodic function, then the steady-state solution will be a superposition of harmonic oscillations with frequencies equal to that of $r(t)$ and integer multiples of these frequencies. And if one of these frequencies is close to the (practical) resonant frequency of the vibrating system (see Sec. 2.8), then the corresponding oscillation may be the dominant part of the response of the system to the external force. This is what the use of Fourier series will show us. Of course, this is quite surprising to an observer unfamiliar with Fourier series, which are highly important in the study of vibrating systems and resonance. Let us discuss the entire situation in terms of a typical example.

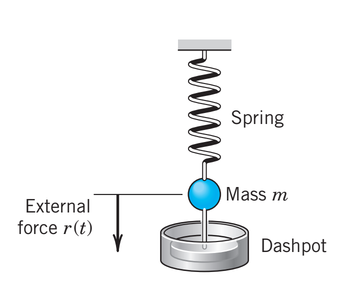


Fig. 274. Vibrating system under consideration

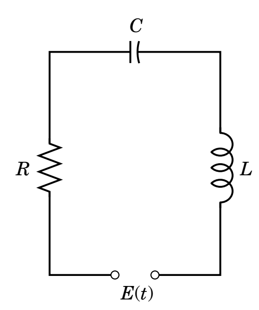


Fig. 275. Electrical analog of the system in Fig. 274 (*RLC-circuit*)

EXAMPLE 1 Forced Oscillations under a Nonsinusoidal Periodic Driving Force

In (1), let $m = 1$ (g), $c = 0.05$ (g/sec), and $k = 25$ (g/sec^2), so that (1) becomes

(2) $$y'' + 0.05y' + 25y = r(t)$$

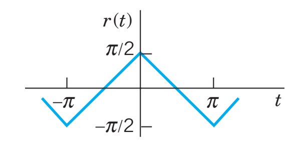


Fig. 276. Force in Example 1

where $r(t)$ is measured in $\text{g} \cdot \text{cm/sec}^2$. Let (Fig. 276)

$$r(t) = \begin{cases} t + \dfrac{\pi}{2} & \text{if} \quad -\pi < t < 0, \\ -t + \dfrac{\pi}{2} & \text{if} \quad 0 < t < \pi, \end{cases} \qquad r(t + 2\pi) = r(t).$$

Find the steady-state solution $y(t)$.

Solution. We represent $r(t)$ by a Fourier series, finding

$$r(t) = \frac{4}{\pi}\left(\cos t + \frac{1}{3^2}\cos 3t + \frac{1}{5^2}\cos 5t + \cdots\right). \tag{3}$$

Then we consider the ODE

$$y'' + 0.05y' + 25y = \frac{4}{n^2\pi}\cos nt \qquad (n = 1, 3, \cdots) \tag{4}$$

whose right side is a single term of the series (3). From Sec. 2.8 we know that the steady-state solution $y_n(t)$ of (4) is of the form

$$y_n = A_n \cos nt + B_n \sin nt. \tag{5}$$

By substituting this into (4) we find that

$$A_n = \frac{4(25 - n^2)}{n^2\pi D_n}, \qquad B_n = \frac{0.2}{n\pi D_n}, \qquad \text{where} \qquad D_n = (25 - n^2)^2 + (0.05n)^2. \tag{6}$$

Since the ODE (2) is linear, we may expect the steady-state solution to be

$$y = y_1 + y_3 + y_5 + \cdots \tag{7}$$

where y_n is given by (5) and (6). In fact, this follows readily by substituting (7) into (2) and using the Fourier series of $r(t)$, provided that termwise differentiation of (7) is permissible. (Readers already familiar with the notion of uniform convergence [Sec. 15.5] may prove that (7) may be differentiated term by term.)

From (6) we find that the amplitude of (5) is (a factor $\sqrt{D_n}$ cancels out)

$$C_n = \sqrt{A_n^2 + B_n^2} = \frac{4}{n^2\pi\sqrt{D_n}}.$$

Values of the first few amplitudes are

$$C_1 = 0.0531 \quad C_3 = 0.0088 \quad C_5 = 0.2037 \quad C_7 = 0.0011 \quad C_9 = 0.0003.$$

Figure 277 shows the input (multiplied by 0.1) and the output. For $n = 5$ the quantity D_n is very small, the denominator of C_5 is small, and C_5 is so large that y_5 is the dominating term in (7). Hence the output is almost a harmonic oscillation of five times the frequency of the driving force, a little distorted due to the term y_1, whose amplitude is about 25% of that of y_5. You could make the situation still more extreme by decreasing the damping constant c. Try it. ■

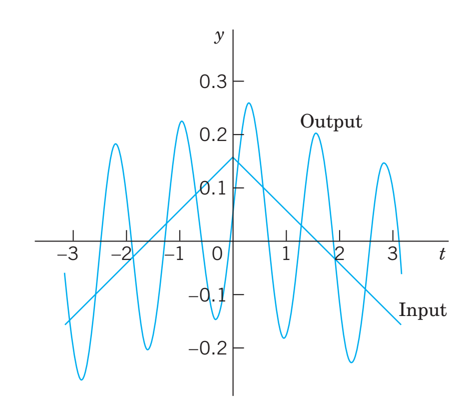


Fig. 277. Input and steady-state output in Example 1

PROBLEM SET 11.3

1. **Coefficients C_n.** Derive the formula for C_n from A_n and B_n.
2. **Change of spring and damping.** In Example 1, what happens to the amplitudes C_n if we take a stiffer spring, say, of $k = 49$? If we increase the damping?
3. **Phase shift.** Explain the role of the B_n's. What happens if we let $c \to 0$?
4. **Differentiation of input.** In Example 1, what happens if we replace $r(t)$ with its derivative, the rectangular wave? What is the ratio of the new C_n to the old ones?
5. **Sign of coefficients.** Some of the A_n in Example 1 are positive, some negative. All B_n are positive. Is this physically understandable?

6–11 GENERAL SOLUTION

Find a general solution of the ODE $y'' + \omega^2 y = r(t)$ with $r(t)$ as given. Show the details of your work.

6. $r(t) = \sin \alpha t + \sin \beta t$, $\omega^2 \neq \alpha^2, \beta^2$
7. $r(t) = \sin t$, $\omega = 0.5, 0.9, 1.1, 1.5, 10$
8. **Rectifier.** $r(t) = \pi/4\,|\cos t|$ if $-\pi < t < \pi$ and $r(t + 2\pi) = r(t)$, $|\omega| \neq 0, 2, 4, \cdots$
9. What kind of solution is excluded in Prob. 8 by $|\omega| \neq 0, 2, 4, \cdots$?
10. **Rectifier.** $r(t) = \pi/4\,|\sin t|$ if $0 < t < 2\pi$ and $r(t + 2\pi) = r(t)$, $|\omega| \neq 0, 2, 4, \cdots$
11. $r(t) = \begin{cases} -1 & \text{if } -\pi < t < 0 \\ 1 & \text{if } 0 < t < \pi, \end{cases}$ $\quad |\omega| \neq 1, 3, 5, \cdots$
12. **CAS Program.** Write a program for solving the ODE just considered and for jointly graphing input and output of an initial value problem involving that ODE. Apply the program to Probs. 7 and 11 with initial values of your choice.

13–16 STEADY-STATE DAMPED OSCILLATIONS

Find the steady-state oscillations of $y'' + cy' + y = r(t)$ with $c > 0$ and $r(t)$ as given. Note that the spring constant is $k = 1$. Show the details. In Probs. 14–16 sketch $r(t)$.

13. $r(t) = \sum_{n=1}^{N} (a_n \cos nt + b_n \sin nt)$
14. $r(t) = \begin{cases} -1 & \text{if } -\pi < t < 0 \\ 1 & \text{if } 0 < t < \pi \end{cases}$ and $r(t + 2\pi) = r(t)$
15. $r(t) = t(\pi^2 - t^2)$ if $-\pi < t < \pi$ and $r(t + 2\pi) = r(t)$
16. $r(t) = \begin{cases} t & \text{if } -\pi/2 < t < \pi/2 \\ \pi - t & \text{if } \pi/2 < t < 3\pi/2 \end{cases}$ and $r(t + 2\pi) = r(t)$

17–19 *RLC*-CIRCUIT

Find the steady-state current $I(t)$ in the *RLC*-circuit in Fig. 275, where $R = 10\ \Omega$, $L = 1$ H, $C = 10^{-1}$ F and with $E(t)$ V as follows and periodic with period 2π. Graph or sketch the first four partial sums. Note that the coefficients of the solution decrease rapidly. *Hint.* Remember that the ODE contains $E'(t)$, not $E(t)$, cf. Sec. 2.9.

17. $E(t) = \begin{cases} -50t^2 & \text{if } -\pi < t < 0 \\ 50t^2 & \text{if } 0 < t < \pi \end{cases}$

18. $E(t) = \begin{cases} 100\,(t - t^2) & \text{if} \quad -\pi < t < 0 \\ 100\,(t + t^2) & \text{if} \quad 0 < t < \pi \end{cases}$

19. $E(t) = 200t(\pi^2 - t^2) \quad (-\pi < t < \pi)$

20. CAS EXPERIMENT. Maximum Output Term. Graph and discuss outputs of $y'' + cy' + ky = r(t)$ with $r(t)$ as in Example 1 for various c and k with emphasis on the maximum C_n and its ratio to the second largest $|C_n|$.

11.4 Approximation by Trigonometric Polynomials

Fourier series play a prominent role not only in differential equations but also in **approximation theory**, an area that is concerned with approximating functions by other functions—usually simpler functions. Here is how Fourier series come into the picture.

Let $f(x)$ be a function on the interval $-\pi \leqq x \leqq \pi$ that can be represented on this interval by a Fourier series. Then the ***N*th partial sum** of the Fourier series

(1) $$f(x) \approx a_0 + \sum_{n=1}^{N} (a_n \cos nx + b_n \sin nx)$$

is an approximation of the given $f(x)$. In (1) we choose an arbitrary N and keep it fixed. Then we ask whether (1) is the "best" approximation of f by a **trigonometric polynomial of the same degree N**, that is, by a function of the form

(2) $$F(x) = A_0 + \sum_{n=1}^{N} (A_n \cos nx + B_n \sin nx) \qquad (N \text{ fixed}).$$

Here, "best" means that the "error" of the approximation is as small as possible.

Of course we must first define what we mean by the **error** of such an approximation. We could choose the maximum of $|f(x) - F(x)|$. But in connection with Fourier series it is better to choose a definition of error that measures the goodness of agreement between f and F *on the whole interval* $-\pi \leqq x \leqq \pi$. This is preferable since the sum f of a Fourier series may have jumps: F in Fig. 278 is a good overall approximation of f, but the maximum of $|f(x) - F(x)|$ (more precisely, the *supremum*) is large. We choose

(3) $$E = \int_{-\pi}^{\pi} (f - F)^2\, dx.$$

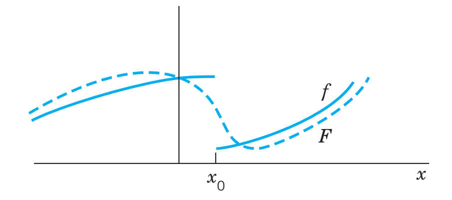


Fig. 278. Error of approximation

This is called the **square error** of F relative to the function f on the interval $-\pi \leqq x \leqq \pi$. Clearly, $E \geqq 0$.

N being fixed, we want to determine the coefficients in (2) such that E is minimum. Since $(f - F)^2 = f^2 - 2fF + F^2$, we have

$$E = \int_{-\pi}^{\pi} f^2\,dx - 2\int_{-\pi}^{\pi} fF\,dx + \int_{-\pi}^{\pi} F^2\,dx. \tag{4}$$

We square (2), insert it into the last integral in (4), and evaluate the occurring integrals. This gives integrals of $\cos^2 nx$ and $\sin^2 nx$ $(n \geqq 1)$, which equal π, and integrals of $\cos nx$, $\sin nx$, and $(\cos nx)(\sin mx)$, which are zero (just as in Sec. 11.1). Thus

$$\int_{-\pi}^{\pi} F^2\,dx = \int_{-\pi}^{\pi} \left[A_0 + \sum_{n=1}^{N} (A_n \cos nx + B_n \sin nx)\right]^2 dx$$
$$= \pi(2A_0^2 + A_1^2 + \cdots + A_N^2 + B_1^2 + \cdots + B_N^2).$$

We now insert (2) into the integral of fF in (4). This gives integrals of $f \cos nx$ as well as $f \sin nx$, just as in Euler's formulas, Sec. 11.1, for a_n and b_n (each multiplied by A_n or B_n). Hence

$$\int_{-\pi}^{\pi} fF\,dx = \pi(2A_0a_0 + A_1a_1 + \cdots + A_Na_N + B_1b_1 + \cdots + B_Nb_N).$$

With these expressions, (4) becomes

$$\begin{aligned} E = \int_{-\pi}^{\pi} f^2\,dx &- 2\pi\left[2A_0a_0 + \sum_{n=1}^{N} (A_na_n + B_nb_n)\right] \\ &+ \pi\left[2A_0^2 + \sum_{n=1}^{N} (A_n^2 + B_n^2)\right]. \end{aligned} \tag{5}$$

We now take $A_n = a_n$ and $B_n = b_n$ in (2). Then in (5) the second line cancels half of the integral-free expression in the first line. Hence for this choice of the coefficients of F the square error, call it E^*, is

$$E^* = \int_{-\pi}^{\pi} f^2\,dx - \pi\left[2a_0^2 + \sum_{n=1}^{N} (a_n^2 + b_n^2)\right]. \tag{6}$$

We finally subtract (6) from (5). Then the integrals drop out and we get terms $A_n^2 - 2A_na_n + a_n^2 = (A_n - an)^2$ and similar terms $(B_n - b_n)^2$:

$$E - E^* = \pi\left\{2(A_0 - a_0)^2 + \sum_{n=1}^{N}[(A_n - a_n)^2 + (B_n - b_n)^2]\right\}.$$

Since the sum of squares of real numbers on the right cannot be negative,

$$E - E^* \geqq 0, \qquad \text{thus} \qquad E \geqq E^*,$$

and $E = E^*$ if and only if $A_0 = a_0, \cdots, B_N = b_N$. This proves the following fundamental minimum property of the partial sums of Fourier series.

THEOREM 1

Minimum Square Error

The square error of F in (2) (with fixed N) relative to f on the interval $-\pi \leqq x \leqq \pi$ is minimum if and only if the coefficients of F in (2) are the Fourier coefficients of f. This minimum value E^* is given by (6).

From (6) we see that E^* cannot increase as N increases, but may decrease. Hence *with increasing N the partial sums of the Fourier series of f yield better and better approximations to f,* considered from the viewpoint of the square error.

Since $E^* \geqq 0$ and (6) holds for every N, we obtain from (6) the important **Bessel's inequality**

$$2a_0^2 + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) \leqq \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\, dx \tag{7}$$

for the Fourier coefficients of any function f for which integral on the right exists. (For F. W. Bessel see Sec. 5.5.)

It can be shown (see [C12] in App. 1) that for such a function f, **Parseval's theorem** holds; that is, formula (7) holds with the equality sign, so that it becomes **Parseval's identity**[3]

$$2a_0^2 + \sum_{n=1}^{\infty} (a_n^2 + b_n^2) = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)^2\, dx. \tag{8}$$

EXAMPLE 1 **Minimum Square Error for the Sawtooth Wave**

Compute the minimum square error E^* of $F(x)$ with $N = 1, 2, \cdots, 10, 20, \cdots, 100$ and 1000 relative to

$$f(x) = x + \pi \qquad (-\pi < x < \pi)$$

on the interval $-\pi \leqq x \leqq \pi$.

Solution. $F(x) = \pi + 2\left(\sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots + \frac{(-1)^{N+1}}{N}\sin Nx\right)$ by Example 3 in Sec. 11.3. From this and (6),

$$E^* = \int_{-\pi}^{\pi} (x + \pi)^2\, dx - \pi\left(2\pi^2 + 4\sum_{n=1}^{N}\frac{1}{n^2}\right).$$

Numeric values are:

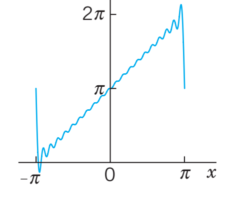


Fig. 279. F with $N = 20$ in Example 1

N	E^*	N	E^*	N	E^*	N	E^*
1	8.1045	6	1.9295	20	0.6129	70	0.1782
2	4.9629	7	1.6730	30	0.4120	80	0.1561
3	3.5666	8	1.4767	40	0.3103	90	0.1389
4	2.7812	9	1.3216	50	0.2488	100	0.1250
5	2.2786	10	1.1959	60	0.2077	1000	0.0126

[3]MARC ANTOINE PARSEVAL (1755–1836), French mathematician. A physical interpretation of the identity follows in the next section.

$F = S_1, S_2, S_3$ are shown in Fig. 269 in Sec. 11.2, and $F = S_{20}$ is shown in Fig. 279. Although $|f(x) - F(x)|$ is large at $\pm\pi$ (how large?), where f is discontinuous, F approximates f quite well on the whole interval, except near $\pm\pi$, where "waves" remain owing to the "Gibbs phenomenon," which we shall discuss in the next section.

Can you think of functions f for which E^* decreases more quickly with increasing N? ■

PROBLEM SET 11.4

1. **CAS Problem.** Do the numeric and graphic work in Example 1 in the text.

2–5 MINIMUM SQUARE ERROR

Find the trigonometric polynomial $F(x)$ of the form (2) for which the square error with respect to the given $f(x)$ on the interval $-\pi < x < \pi$ is minimum. Compute the minimum value for $N = 1, 2, \cdots, 5$ (or also for larger values if you have a CAS).

2. $f(x) = x \quad (-\pi < x < \pi)$
3. $f(x) = |x| \quad (-\pi < x < \pi)$
4. $f(x) = x^2 \quad (-\pi < x < \pi)$
5. $f(x) = \begin{cases} -1 & \text{if} \quad -\pi < x < 0 \\ 1 & \text{if} \quad 0 < x < \pi \end{cases}$
6. Why are the square errors in Prob. 5 substantially larger than in Prob. 3?
7. $f(x) = x^3 \quad (-\pi < x < \pi)$
8. $f(x) = |\sin x| \quad (-\pi < x < \pi)$, full-wave rectifier
9. **Monotonicity.** Show that the minimum square error (6) is a monotone decreasing function of N. How can you use this in practice?
10. **CAS EXPERIMENT. Size and Decrease of E^*.** Compare the size of the minimum square error E^* for functions of your choice. Find experimentally the factors on which the decrease of E^* with N depends. For each function considered find the smallest N such that $E^* < 0.1$.

11–15 PARSEVALS'S IDENTITY

Using (8), prove that the series has the indicated sum. Compute the first few partial sums to see that the convergence is rapid.

11. $1 + \frac{1}{3^2} + \frac{1}{5^2} + \cdots = \frac{\pi^2}{8} = 1.233700550$

    Use Example 1 in Sec. 11.1.
12. $1 + \frac{1}{2^4} + \frac{1}{3^4} + \cdots = \frac{\pi^4}{90} = 1.082323234$

    Use Prob. 14 in Sec. 11.1.
13. $1 + \frac{1}{3^4} + \frac{1}{5^4} + \frac{1}{7^4} + \cdots = \frac{\pi^4}{96} = 1.014678032$

    Use Prob. 17 in Sec. 11.1.
14. $\int_{-\pi}^{\pi} \cos^4 x \, dx = \frac{3\pi}{4}$
15. $\int_{-\pi}^{\pi} \cos^6 x \, dx = \frac{5\pi}{8}$

11.5 Sturm–Liouville Problems. Orthogonal Functions

The idea of the Fourier series was to represent general periodic functions in terms of cosines and sines. The latter formed a *trigonometric system.* This trigonometric system has the desirable property of orthogonality which allows us to compute the coefficient of the Fourier series by the Euler formulas.

The question then arises, can this approach be generalized? That is, can we replace the trigonometric system of Sec. 11.1 by other *orthogonal systems* (*sets of other orthogonal functions*)? The answer is "yes" and will lead to generalized Fourier series, including the Fourier–Legendre series and the Fourier–Bessel series in Sec. 11.6.

To prepare for this generalization, we first have to introduce the concept of a Sturm–Liouville Problem. (The motivation for this approach will become clear as you read on.) Consider a second-order ODE of the form

(1) $$[p(x)y']' + [q(x) + \lambda r(x)]y = 0$$

on some interval $a \leqq x \leqq b$, satisfying conditions of the form

(2)
$$\textbf{(a)} \quad k_1 y + k_2 y' = 0 \quad \text{at } x = a$$
$$\textbf{(b)} \quad l_1 y + l_2 y' = 0 \quad \text{at } x = b.$$

Here λ is a parameter, and k_1, k_2, l_1, l_2 are given real constants. Furthermore, at least one of each constant in each condition (2) must be different from zero. (We will see in Example 1 that, if $p(x) = r(x) = 1$ and $q(x) = 0$, then $\sin \sqrt{\lambda}x$ and $\cos \sqrt{\lambda}x$ satisfy (1) and constants can be found to satisfy (2).) Equation (1) is known as a **Sturm–Liouville equation**.[4] Together with conditions 2(a), 2(b) it is know as the **Sturm–Liouville problem**. It is an example of a boundary value problem.

A **boundary value problem** consists of an ODE and given boundary conditions referring to the two boundary points (endpoints) $x = a$ and $x = b$ of a given interval $a \leqq x \leqq b$.

The goal is to solve these type of problems. To do so, we have to consider

Eigenvalues, Eigenfunctions

Clearly, $y \equiv 0$ is a solution—the "**trivial solution**"—of the problem (1), (2) for any λ because (1) is homogeneous and (2) has zeros on the right. This is of no interest. We want to find **eigenfunctions** $y(x)$, that is, solutions of (1) satisfying (2) without being identically zero. We call a number λ for which an eigenfunction exists an **eigenvalue** of the Sturm–Liouville problem (1), (2).

Many important ODEs in engineering can be written as Sturm–Liouville equations. The following example serves as a case in point.

EXAMPLE 1 Trigonometric Functions as Eigenfunctions. Vibrating String

Find the eigenvalues and eigenfunctions of the Sturm–Liouville problem

(3) $$y'' + \lambda y = 0, \qquad y(0) = 0, \quad y(\pi) = 0.$$

This problem arises, for instance, if an elastic string (a violin string, for example) is stretched a little and fixed at its ends $x = 0$ and $x = \pi$ and then allowed to vibrate. Then $y(x)$ is the "space function" of the deflection $u(x, t)$ of the string, assumed in the form $u(x, t) = y(x)w(t)$, where t is time. (This model will be discussed in great detail in Secs, 12.2–12.4.)

Solution. From (1) nad (2) we see that $p = 1, q = 0, r = 1$ in (1), and $a = 0, b = \pi, k_1 = l_1 = 1$, $k_2 = l_2 = 0$ in (2). For negative $\lambda = -\nu^2$ a general solution of the ODE in (3) is $y(x) = c_1 e^{\nu x} + c_2 e^{-\nu x}$. From the boundary conditions we obtain $c_1 = c_2 = 0$, so that $y \equiv 0$, which is not an eigenfunction. For $\lambda = 0$ the situation is similar. For positive $\lambda = \nu^2$ a general solution is

$$y(x) = A \cos \nu x + B \sin \nu x.$$

[4]JACQUES CHARLES FRANÇOIS STURM (1803–1855) was born and studied in Switzerland and then moved to Paris, where he later became the successor of Poisson in the chair of mechanics at the Sorbonne (the University of Paris).

JOSEPH LIOUVILLE (1809–1882), French mathematician and professor in Paris, contributed to various fields in mathematics and is particularly known by his important work in complex analysis (Liouville's theorem; Sec. 14.4), special functions, differential geometry, and number theory.

From the first boundary condition we obtain $y(0) = A = 0$. The second boundary condition then yields

$$y(\pi) = B \sin \nu\pi = 0, \qquad \text{thus} \qquad \nu = 0, \pm 1, \pm 2, \cdots.$$

For $\nu = 0$ we have $y \equiv 0$. For $\lambda = \nu^2 = 1, 4, 9, 16, \cdots$, taking $B = 1$, we obtain

$$y(x) = \sin \nu x \qquad (\nu = \sqrt{\lambda} = 1, 2, \cdots).$$

Hence the eigenvalues of the problem are $\lambda = \nu^2$, where $\nu = 1, 2, \cdots$, and corresponding eigenfunctions are $y(x) = \sin \nu x$, where $\nu = 1, 2 \cdots$. ■

Note that the solution to this problem is precisely the trigonometric system of the Fourier series considered earlier. It can be shown that, under rather general conditions on the functions p, q, r in (1), the Sturm–Liouville problem (1), (2) has infinitely many eigenvalues. The corresponding rather complicated theory can be found in Ref. [All] listed in App. 1.

Furthermore, if p, q, r, and p' in (1) are real-valued and continuous on the interval $a \leqq x \leqq b$ and r is positive throughout that interval (or negative throughout that interval), then all the eigenvalues of the Sturm–Liouville problem (1), (2) are real. (Proof in App. 4.) This is what the engineer would expect since eigenvalues are often related to frequencies, energies, or other physical quantities that must be real.

The most remarkable and important property of eigenfunctions of Sturm–Liouville problems is their ***orthogonality***, which will be crucial in series developments in terms of eigenfunctions, as we shall see in the next section. This suggests that we should next consider orthogonal functions.

Orthogonal Functions

Functions $y_1(x), y_2(x), \cdots$ defined on some interval $a \leqq x \leqq b$ are called **orthogonal** on this interval with respect to the **weight function** $r(x) > 0$ if for all m and all n different from m,

$$(y_m, y_n) = \int_a^b r(x) y_m(x) y_n(x) \, dx = 0 \qquad (m \neq n). \tag{4}$$

(y_m, y_n) is a ***standard notation*** for this integral. **The norm** $\|y_m\|$ of y_m is defined by

$$\|y_m\| = \sqrt{(y_m, y_m)} = \sqrt{\int_a^b r(x) y_m^2(x) \, dx}. \tag{5}$$

Note that this is the square root of the integral in (4) with $n = m$.

The functions $y_1, y_2, \cdots$ are called **orthonormal** on $a \leqq x \leqq b$ if they are orthogonal on this interval and all have norm 1. Then we can write (4), (5) jointly by using the **Kronecker symbol**[5] δ_{mn}, namely,

$$(y_m, y_n) = \int_a^b r(x) y_m(x) y_n(x) \, dx = \delta_{mn} = \begin{cases} 0 & \text{if} \quad m \neq \boldsymbol{n} \\ 1 & \text{if} \quad m = \boldsymbol{n}. \end{cases}$$

[5] LEOPOLD KRONECKER (1823–1891). German mathematician at Berlin University, who made important contributions to algebra, group theory, and number theory.

If $r(x) = 1$, we more briefly call the functions *orthogonal* instead of orthogonal with respect to $r(x) = 1$; similarly for orthognormality. Then

$$(y_m, y_n) = \int_a^b y_m(x) y_n(x)\, dx = 0 \quad (m \neq n), \qquad \|y_m\| = \sqrt{(y_m, y_n)} = \sqrt{\int_a^b y_m^2(x)\, dx}.$$

The next example serves as an illustration of the material on orthogonal functions just discussed.

EXAMPLE 2 **Orthogonal Functions. Orthonormal Functions. Notation**

The functions $y_m(x) = \sin mx$, $m = 1, 2, \cdots$ form an orthogonal set on the interval $-\pi \leqq x \leqq \pi$, because for $m \neq n$ we obtain by integration [see (11) in App. A3.1]

$$(y_m, y_n) = \int_{-\pi}^{\pi} \sin mx \sin nx\, dx = \frac{1}{2}\int_{-\pi}^{\pi} \cos (m - n)x\, dx - \frac{1}{2}\int_{-\pi}^{\pi} \cos (m + n)x\, dx = 0, \quad (m \neq n).$$

The norm $\| y_m \| = \sqrt{(y_m, y_m)}$ equals $\sqrt{\pi}$ because

$$\|y_m\|^2 = (y_m, y_m) = \int_{-\pi}^{\pi} \sin^2 mx\, dx = \pi \qquad (m = 1, 2, \cdots)$$

Hence the corresponding orthonormal set, obtained by division by the norm, is

$$\frac{\sin x}{\sqrt{\pi}}, \quad \frac{\sin 2x}{\sqrt{\pi}}, \quad \frac{\sin 3x}{\sqrt{\pi}}, \quad \cdots.$$ ■

Theorem 1 shows that for any Sturm–Liouville problem, the eigenfunctions associated with these problems are orthogonal. This means, in practice, if we can formulate a problem as a Sturm–Liouville problem, then by this theorem we are guaranteed orthogonality.

THEOREM 1 **Orthogonality of Eigenfunctions of Sturm–Liouville Problems**

Suppose that the functions p, q, r, and p' in the Sturm–Liouville equation (1) *are real-valued and continuous and $r(x) > 0$ on the interval $a \leqq x \leqq b$. Let $y_m(x)$ and $y_n(x)$ be eigenfunctions of the Sturm–Liouville problem* (1), (2) *that correspond to different eigenvalues λ_m and λ_n, respectively. Then y_m, y_n are orthogonal on that interval with respect to the weight function r, that is,*

$$(y_m, y_n) = \int_a^b r(x) y_m(x) y_n(x)\, dx = 0 \qquad (m \neq n). \tag{6}$$

If $p(a) = 0$, then (2a) *can be dropped from the problem. If $p(b) = 0$, then* (2b) can be dropped. [It is then required that y and y' remain bounded at such a point, and the problem is called **singular**, as opposed to a **regular problem** in which (2) is used.]

If $p(a) = p(b)$, then (2) *can be replaced by the* "**periodic boundary conditions**"

$$y(a) = y(b), \qquad y'(a) = y'(b). \tag{7}$$

The boundary value problem consisting of the Sturm–Liouville equation (1) and the periodic boundary conditions (7) is called a **periodic Sturm–Liouville problem**.

PROOF By assumption, y_m and y_n satisfy the Sturm–Liouville equations

$$(py_m')' + (q + \lambda_m r)y_m = 0$$
$$(py_n')' + (q + \lambda_n r)y_n = 0$$

respectively. We multiply the first equation by y_n, the second by $-y_m$, and add,

$$(\lambda_m - \lambda_n)ry_m y_n = y_m(py_n')' - y_n(py_m')' = [(py_n')y_m - [(py_m')y_n]'$$

where the last equality can be readily verified by performing the indicated differentiation of the last expression in brackets. This expression is continuous on $a \leqq x \leqq b$ since p and p' are continuous by assumption and y_m, y_n are solutions of (1). Integrating over x from a to b, we thus obtain

$$(8) \qquad (\lambda_m - \lambda_n)\int_a^b ry_m y_n \, dx = [p(y_n' y_m - y_m' y_n)]_a^b \qquad (a < b).$$

The expression on the right equals the sum of the subsequent Lines 1 and 2,

$$(9) \qquad \begin{aligned} &p(b)[y_n'(b)y_m(b) - y_m'(b)y_n(b)] && \text{(Line 1)} \\ -&p(a)[y_n'(a)y_m(a) - y_m'(a)y_n(a)] && \text{(Line 2).} \end{aligned}$$

Hence if (9) is zero, (8) with $\lambda_m - \lambda_n \neq 0$ implies the orthogonality (6). Accordingly, we have to show that (9) is zero, using the boundary conditions (2) as needed.

Case 1. $\boldsymbol{p(a) = p(b) = 0.}$ Clearly, (9) is zero, and (2) is not needed.
Case 2. $\boldsymbol{p(a) \neq 0, p(b) = 0.}$ Line 1 of (9) is zero. Consider Line 2. From (2a) we have

$$k_1 y_n(a) + k_2 y_n'(a) = 0,$$
$$k_1 y_m(a) + k_2 y_m'(a) = 0.$$

Let $k_2 \neq 0$. We multiply the first equation by $y_m(a)$, the last by $-y_n(a)$ and add,

$$k_2[y_n'(a)y_m(a) - y_m'(a)y_n(a)] = 0.$$

This is k_2 times Line 2 of (9), which thus is zero since $k_2 \neq 0$. If $k_2 = 0$, then $k_1 \neq 0$ by assumption, and the argument of proof is similar.

Case 3. $\boldsymbol{p(a) = 0, p(b) \neq 0.}$ Line 2 of (9) is zero. From (2b) it follows that Line 1 of (9) is zero; this is similar to Case 2.
Case 4. $\boldsymbol{p(a) \neq 0, p(b) \neq 0.}$ We use both (2a) and (2b) and proceed as in Cases 2 and 3.
Case 5. $\boldsymbol{p(a) = p(b).}$ Then (9) becomes

$$p(b)[y_n'(b)y_m(b) - y_m'(b)y_n(b) - y_n'(a)y_m(a) + y_m'(a)y_n(a)].$$

The expression in brackets $[\cdots]$ is zero, either by (2) used as before, or more directly by (7). Hence in this case, (7) can be used instead of (2), as claimed. This completes the proof of Theorem 1. ■

EXAMPLE 3 **Application of Theorem 1. Vibrating String**

The ODE in Example 1 is a Sturm–Liouville equation with $p = 1$, $q = 0$, and $r = 1$. From Theorem 1 it follows that the eigenfunctions $y_m = \sin mx$ $(m = 1, 2, \cdots)$ are orthogonal on the interval $0 \leqq x \leqq \pi$. ■

Example 3 confirms, from this new perspective, that the trigonometric system underlying the Fourier series is orthogonal, as we knew from Sec. 11.1.

EXAMPLE 4 Application of Theorem 1. Orthogonlity of the Legendre Polynomials

Legendre's equation $(1 - x^2)y'' - 2xy' + n(n + 1)y = 0$ may be written

$$[(1 - x^2)y']' + \lambda y = 0 \qquad \lambda = n(n + 1).$$

Hence, this is a Sturm–Liouville equation (1) with $p = 1 - x^2$, $q = 0$, and $r = 1$. Since $p(-1) = p(1) = 0$, we need no boundary conditions, but have a "***singular*** ***Sturm–Liouville problem***" on the interval $-1 \leqq x \leqq 1$. We know that for $n = 0, 1, \cdots$, hence $\lambda = 0, 1 \cdot 2, 2 \cdot 3, \cdots$, the Legendre polynomials $P_n(x)$ are solutions of the problem. Hence these are the eigenfunctions. From Theorem 1 it follows that they are orthogonal on that interval, that is,

$$\int_{-1}^{1} P_m(x)P_n(x)\, dx = 0 \qquad (m \neq n). \tag{10}$$

What we have seen is that the trigonometric system, underlying the Fourier series, is a solution to a Sturm–Liouville problem, as shown in Example 1, and that this trigonometric system is orthogonal, which we knew from Sec. 11.1 and confirmed in Example 3.

PROBLEM SET 11.5

1. **Proof of Theorem 1.** Carry out the details in Cases 3 and 4.

2–6 ORTHOGONALITY

2. **Normalization of eigenfunctions** y_m of (1), (2) means that we multiply y_m by a nonzero constant c_m such that $c_m y_m$ has norm 1. Show that $z_m = cy_m$ with *any* $c \neq 0$ is an eigenfunction for the eigenvalue corresponding to y_m.
3. **Change of x.** Show that if the functions $y_0(x), y_1(x), \cdots$ form an orthogonal set on an interval $a \leqq x \leqq b$ (with $r(x) = 1$), then the functions $y_0(ct + k), y_1(ct + k), \cdots, c > 0$, form an orthogonal set on the interval $(a - k)/c \leqq t \leqq (b - k)/c$.
4. **Change of x.** Using Prob. 3, derive the orthogonality of 1, $\cos \pi x$, $\sin \pi x$, $\cos 2\pi x$, $\sin 2\pi x$, $\cdots$ on $-1 \leqq x \leqq 1$ $(r(x) = 1)$ from that of 1, $\cos x$, $\sin x$, $\cos 2x$, $\sin 2x$, $\cdots$ on $-\pi \leqq x \leqq \pi$.
5. **Legendre polynomials.** Show that the functions $P_n(\cos \theta)$, $n = 0, 1, \cdots$, from an orthogonal set on the interval $0 \leqq \theta \leqq \pi$ with respect to the weight function $\sin \theta$.
6. **Tranformation to Sturm–Liouville form.** Show that $y'' + fy' + (g + \lambda h)\, y = 0$ takes the form (1) if you set $p = \exp(\int f\, dx)$, $q = pg$, $r = hp$. Why would you do such a transformation?

7–15 STURM–LIOUVILLE PROBLEMS

Find the eigenvalues and eigenfunctions. Verify orthogonality. Start by writing the ODE in the form (1), using Prob. 6. Show details of your work.

7. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y(10) = 0$
8. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y(L) = 0$
9. $y'' + \lambda y = 0, \quad y(0) = 0, \quad y'(L) = 0$
10. $y'' + \lambda y = 0, \quad y(0) = y(1), \quad y'(0) = y'(1)$
11. $(y'/x)' + (\lambda + 1)y/x^3 = 0, \quad y(1) = 0, \quad y(e^{\pi}) = 0.$ (Set $x = e^t$.)
12. $y'' - 2y' + (\lambda + 1)y = 0, \quad y(0) = 0, \quad y(1) = 0$
13. $y'' + 8y' + (\lambda + 16)y = 0, \quad y(0) = 0, \quad y(\pi) = 0$
14. **TEAM PROJECT. Special Functions. Orthogonal polynomials** play a great role in applications. For this reason, Legendre polynomials and various other orthogonal polynomials have been studied extensively; see Refs. [GenRef1], [GenRef10] in App. 1. Consider some of the most important ones as follows.

(a) Chebyshev polynomials[6] of the first and second kind are defined by

$$T_n(x) = \cos(n \arccos x)$$
$$U_n(x) = \frac{\sin[(n+1)\arccos x]}{\sqrt{1-x^2}}$$

respectively, where $n = 0, 1, \cdots$. Show that

$$T_0 = 1, \qquad T_1(x) = x, \qquad T_2(x) = 2x^2 - 1,$$
$$T_3(x) = 4x^3 - 3x,$$
$$U_0 = 1, \qquad U_1(x) = 2x, \qquad U_2(x) = 4x^2 - 1,$$
$$U_3(x) = 8x^3 - 4x.$$

Show that the Chebyshev polynomials $T_n(x)$ are orthogonal on the interval $-1 \leqq x \leqq 1$ with respect to the weight function $r(x) = 1/\sqrt{1-x^2}$. (*Hint.* To evaluate the integral, set $\arccos x = \theta$.) Verify that $T_n(x)$, $n = 0, 1, 2, 3$, satisfy the **Chebyshev equation**

$$(1-x^2)y'' - xy' + n^2y = 0.$$

(b) Orthogonality on an infinite interval: Laguerre polynomials[7] are defined by $L_0 = 1$, and

$$L_n(x) = \frac{e^x}{n!}\frac{d^n(x^n e^{-x})}{dx^n}, \qquad n = 1, 2, \cdots.$$

Show that

$$L_n(x) = 1 - x, \qquad L_2(x) = 1 - 2x + x^2/2,$$
$$L_3(x) = 1 - 3x + 3x^2/2 - x^3/6.$$

Prove that the Laguerre polynomials are orthogonal on the positive axis $0 \leqq x < \infty$ with respect to the weight function $r(x) = e^{-x}$. *Hint.* Since the highest power in L_m is x^m, it suffices to show that $\int e^{-x}x^k L_n\, dx = 0$ for $k < n$. Do this by k integrations by parts.

11.6 Orthogonal Series. Generalized Fourier Series

Fourier series are made up of the trigonometric system (Sec. 11.1), which is orthogonal, and orthogonality was essential in obtaining the Euler formulas for the Fourier coefficients. Orthogonality will also give us coefficient formulas for the desired generalized Fourier series, including the Fourier–Legendre series and the Fourier–Bessel series. This generalization is as follows.

Let $y_0, y_1, y_2, \cdots$ be orthogonal with respect to a weight function $r(x)$ on an interval $a \leqq x \leqq b$, and let $f(x)$ be a function that can be represented by a convergent series

(1)
$$f(x) = \sum_{m=0}^{\infty} a_m y_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots.$$

This is called an **orthogonal series**, **orthogonal expansion**, or **generalized Fourier series**. If the y_m are the eigenfunctions of a Sturm–Liouville problem, we call (1) an **eigenfunction expansion**. In (1) we use again m for summation since n will be used as a fixed order of Bessel functions.

Given $f(x)$, we have to determine the coefficients in (1), called the **Fourier constants** *of $f(x)$ with respect to $y_0, y_1, \cdots$*. Because of the orthogonality, this is simple. Similarly to Sec. 11.1, we multiply both sides of (1) by $r(x)y_n(x)$ (n ***fixed***) and then integrate on

[6]PAFNUTI CHEBYSHEV (1821–1894), Russian mathematician, is known for his work in approximation theory and the theory of numbers. Another transliteration of the name is TCHEBICHEF.

[7]EDMOND LAGUERRE (1834–1886), French mathematician, who did research work in geometry and in the theory of infinite series.

both sides from a to b. We assume that term-by-term integration is permissible. (This is justified, for instance, in the case of "uniform convergence," as is shown in Sec. 15.5.) Then we obtain

$$(f, y_n) = \int_a^b r f y_n \, dx = \int_a^b r\left(\sum_{m=0}^{\infty} a_m y_m\right) y_n \, dx = \sum_{m=0}^{\infty} a_m \int_a^b r y_m y_n \, dx = \sum_{m=0}^{\infty} a_m (y_m, y_n).$$

Because of the orthogonality all the integrals on the right are zero, except when $m = n$. Hence the whole infinite series reduces to the single term

$$a_n (y_n, y_n) = a_n \|y_n\|^2. \qquad \text{Thus} \qquad (f, y_n) = a_n \|y_n\|^2.$$

Assuming that all the functions y_n have nonzero norm, we can divide by $\|y_n\|^2$; writing again m for n, to be in agreement with (1), we get the desired formula for the Fourier constants

(2)
$$a_m = \frac{(f, y_m)}{\|y_m\|^2} = \frac{1}{\|y_m\|^2} \int_a^b r(x) f(x) y_m(x) \, dx \qquad (n = 0, 1, \cdots).$$

This formula generalizes the Euler formulas (6) in Sec. 11.1 as well as the principle of their derivation, namely, by orthogonality.

EXAMPLE 1 Fourier–Legendre Series

A **Fourier–Legendre series** is an eigenfunction expansion

$$f(x) = \sum_{m=0}^{\infty} a_m P_m(x) = a_0 P_0 + a_1 P_1(x) + a_2 P_2(x) + \cdots = a_0 + a_1 x + a_2 \left(\tfrac{3}{2} x^2 - \tfrac{1}{2}\right) + \cdots$$

in terms of Legendre polynomials (Sec. 5.3). The latter are the eigenfunctions of the Sturm–Liouville problem in Example 4 of Sec. 11.5 on the interval $-1 \leqq x \leqq 1$. We have $r(x) = 1$ for Legendre's equation, and (2) gives

(3)
$$a_m = \frac{2m + 1}{2} \int_{-1}^{1} f(x) P_m(x) \, dx, \qquad m = 0, 1, \cdots$$

because the norm is

(4)
$$\|P_m\| = \sqrt{\int_{-1}^{1} P_m(x)^2 \, dx} = \sqrt{\frac{2}{2m + 1}} \qquad (m = 0, 1, \cdots)$$

as we state without proof. The proof of (4) is tricky; it uses Rodrigues's formula in Problem Set 5.2 and a reduction of the resulting integral to a quotient of gamma functions.

For instance, let $f(x) = \sin \pi x$. Then we obtain the coefficients

$$a_m = \frac{2m + 1}{2} \int_{-1}^{1} (\sin \pi x) P_m(x) \, dx, \qquad \text{thus} \qquad a_1 = \frac{3}{2} \int_{-1}^{1} x \sin \pi x \, dx = \frac{3}{\pi} = 0.95493, \qquad \text{etc.}$$

Hence the Fourier–Legendre series of $\sin \pi x$ is

$$\sin \pi x = 0.95493P_1(x) - 1.15824P_3(x) + 0.21929P_5(x) - 0.01664P_7(x) + 0.00068P_9(x) - 0.00002P_{11}(x) + \cdots.$$

The coefficient of P_{13} is about $3 \cdot 10^{-7}$. The sum of the first three nonzero terms gives a curve that practically coincides with the sine curve. Can you see why the even-numbered coefficients are zero? Why a_3 is the absolutely biggest coefficient?

EXAMPLE 2 **Fourier–Bessel Series**

These series model vibrating membranes (Sec. 12.9) and other physical systems of circular symmetry. We derive these series in three steps.

***Step 1.* Bessel's equation as a Sturm–Liouville equation.** The Bessel function $J_n(x)$ with fixed integer $n \geqq 0$ satisfies Bessel's equation (Sec. 5.5)

$$\tilde{x}^2 \ddot{J}_n(\tilde{x}) + \tilde{x}\dot{J}_n(\tilde{x}) + (\tilde{x}^2 - n^2)J_n(\tilde{x}) = 0$$

where $\dot{J}_n = dJ_n/d\tilde{x}$ and $\ddot{J}_n = d^2J_n/d\tilde{x}^2$. We set $\tilde{x} = kx$. Then $x = \tilde{x}/k$ and by the chain rule, $\dot{J}_n = dJ_n/d\tilde{x} = (dJ_n/dx)/k$ and $\ddot{J}_n = J_n''/k^2$. In the first two terms of Bessel's equation, k^2 and k drop out and we obtain

$$x^2J_n''(kx) + xJ_n'(kx) + (k^2x^2 - n^2)J_n(kx) = 0.$$

Dividing by x and using $(xJ_n'(kx))' = xJ_n''(kx) + J_n'(kx)$ gives the Sturm–Liouville equation

$$[xJ_n'(kx)]' + \left(-\frac{n^2}{x} + \lambda x\right)J_n(kx) = 0 \qquad \lambda = k^2 \tag{5}$$

with $p(x) = x$, $q(x) = -n^2/x$, $r(x) = x$, and parameter $\lambda = k^2$. Since $p(0) = 0$, Theorem 1 in Sec. 11.5 implies orthogonality on an interval $0 \leqq x \leqq R$ (R given, fixed) of those solutions $J_n(kx)$ that are zero at $x = R$, that is,

$$J_n(kR) = 0 \qquad (n \text{ fixed}). \tag{6}$$

Note that $q(x) = -n^2/x$ is discontinuous at 0, but this does not affect the proof of Theorem 1.

***Step 2.* Orthogonality.** It can be shown (see Ref. [A13]) that $J_n(\tilde{x})$ has infinitely many zeros, say, $\tilde{x} = \alpha_{n,1} < \alpha_{n,2} < \cdots$ (see Fig. 110 in Sec. 5.4 for $n = 0$ and 1). Hence we must have

$$kR = \alpha_{n,m} \qquad \text{thus} \qquad k_{n,m} = \alpha_{n,m}/R \qquad (m = 1, 2, \cdots). \tag{7}$$

This proves the following orthogonality property.

THEOREM 1 **Orthogonality of Bessel Functions**

For each fixed nonnegative integer n the sequence of Bessel functions of the first kind $J_n(k_{n,1}x)$, $J_n(k_{n,2}x)$, $\cdots$ with $k_{n,m}$ as in (7) forms an orthogonal set on the interval $0 \leqq x \leqq R$ with respect to the weight function $r(x) = x$, that is,

$$\int_0^R xJ_n(k_{n,m}x)J_n(k_{n,j}x)\,dx = 0 \qquad (j \neq m,\ n \text{ fixed}). \tag{8}$$

Hence we have obtained ***infinitely many orthogonal sets*** of Bessel functions, one for each of $J_0, J_1, J_2, \cdots$. Each set is orthogonal on an interval $0 \leqq x \leqq R$ with a fixed positive R of our choice and with respect to the weight x. The orthogonal set for J_n is $J_n(k_{n,1}x), J_n(k_{n,2}x), J_n(k_{n,3}x), \cdots$, where n is *fixed* and $k_{n,m}$ is given by (7).

Step 3. **Fourier–Bessel series.** The Fourier–Bessel series corresponding to J_n (n fixed) is

$$f(x) = \sum_{m=1}^{\infty} a_m J_n(k_{n,m}x) = a_1 J_n(k_{n,1}x) + a_2 J_n(k_{n,2}x) + a_3 J_n(k_{n,3}x) + \cdots \quad (n \text{ fixed}). \tag{9}$$

The coefficients are (with $\alpha_{n,m} = k_{n,m}R$)

$$a_m = \frac{2}{R^2 J_{n+1}^2(\alpha_{n,m})} \int_0^R x\, f(x)\, J_n(k_{n,m}x)\, dx, \qquad m = 1, 2, \cdots \tag{10}$$

because the square of the norm is

$$\|J_n(k_{n,m}x)\|^2 = \int_0^R x J_n^2(k_{n,m}x)\, dx = \frac{R^2}{2} J_{n+1}^2(k_{n,m}R) \tag{11}$$

as we state without proof (which is tricky; see the discussion beginning on p. 576 of [A13]). ■

EXAMPLE 3 Special Fourier–Bessel Series

For instance, let us consider $f(x) = 1 - x^2$ and take $R = 1$ and $n = 0$ in the series (9), simply writing λ for $\alpha_{0,m}$. Then $k_{n,m} = \alpha_{0,m} = \lambda = 2.405, 5.520, 8.654, 11.792$, etc. (use a CAS or Table A1 in App. 5). Next we calculate the coefficients a_m by (10)

$$a_m = \frac{2}{J_1^2(\lambda)} \int_0^1 x(1 - x^2) J_0(\lambda x)\, dx.$$

This can be integrated by a CAS or by formulas as follows. First use $[xJ_1(\lambda x)]' = \lambda x J_0(\lambda x)$ from Theorem 1 in Sec. 5.4 and then integration by parts,

$$a_m = \frac{2}{J_1^2(\lambda)} \int_0^1 x(1 - x^2) J_0(\lambda x)\, dx = \frac{2}{J_1^2(\lambda)} \left[\frac{1}{\lambda}(1 - x^2) x J_1(\lambda x) \Big|_0^1 - \frac{1}{\lambda} \int_0^1 x J_1(\lambda x)(-2x)\, dx \right].$$

The integral-free part is zero. The remaining integral can be evaluated by $[x^2 J_2(\lambda x)]' = \lambda x^2 J_1(\lambda x)$ from Theorem 1 in Sec. 5.4. This gives

$$a_m = \frac{4 J_2(\lambda)}{\lambda^2 J_1^2(\lambda)} \qquad (\lambda = \alpha_{0,m}).$$

Numeric values can be obtained from a CAS (or from the table on p. 409 of Ref. [GenRef1] in App. 1, together with the formula $J_2 = 2x^{-1}J_1 - J_0$ in Theorem 1 of Sec. 5.4). This gives the eigenfunction expansion of $1 - x^2$ in terms of Bessel functions J_0, that is,

$$1 - x^2 = 1.1081 J_0(2.405x) - 0.1398 J_0(5.520x) + 0.0455 J_0(8.654x) - 0.0210 J_0(11.792x) + \cdots.$$

A graph would show that the curve of $1 - x^2$ and that of the sum of first three terms practically coincide. ■

Mean Square Convergence. Completeness

Ideas on approximation in the last section generalize from Fourier series to orthogonal series (1) that are made up of an orthonormal set that is "complete," that is, consists of "sufficiently many" functions so that (1) can represent large classes of other functions (definition below).

In this connection, convergence is **convergence in the norm**, also called **mean-square convergence**; that is, a sequence of functions f_k is called **convergent** *with the limit f* if

$$\lim_{k \to \infty} \|f_k - f\| = 0; \tag{12*}$$

written out by (5) in Sec. 11.5 (where we can drop the square root, as this does not affect the limit)

(12) $$\lim_{k\to\infty}\int_a^b r(x)[f_k(x) - f(x)]^2\,dx = 0.$$

Accordingly, the series (1) converges and represents f if

(13) $$\lim_{k\to\infty}\int_a^b r(x)[s_k(x) - f(x)]^2\,dx = 0$$

where s_k is the kth partial sum of (1).

(14) $$s_k(x) = \sum_{m=0}^{k} a_m y_m(x).$$

Note that the integral in (13) generalizes (3) in Sec. 11.4.

We now define completeness. An **orthonormal** set $y_0, y_1, \cdots$ on an interval $a \leqq x \leqq b$ is **complete** *in a set of functions* S defined on $a \leqq x \leqq b$ if we can approximate every f belonging to S arbitrarily closely in the norm by a linear combination $a_0y_0 + a_1y_1 + \cdots + a_ky_k$, that is, technically, if for every $\epsilon > 0$ we can find constants $a_0, \cdots, a_k$ (with k large enough) such that

(15) $$\|f - (a_0y_0 + \cdots + a_ky_k)\| < \epsilon.$$

Ref. [GenRef7] in App. 1 uses the more modern term **total** for *complete*.

We can now extend the ideas in Sec. 11.4 that guided us from (3) in Sec. 11.4 to Bessel's and Parseval's formulas (7) and (8) in that section. Performing the square in (13) and using (14), we first have (analog of (4) in Sec. 11.4)

$$\int_a^b r(x)[s_k(x) - f(x)]^2\,dx = \int_a^b rs_k^2\,dx - 2\int_a^b rfs_k\,dx + \int_a^b rf^2\,dx$$

$$= \int_a^b r\left[\sum_{m=0}^{k} a_m y_m\right]^2 dx - 2\sum_{m=0}^{k} a_m\int_a^b rfy_m\,dx + \int_a^b rf^2\,dx.$$

The first integral on the right equals $\sum a_m^2$ because $\int ry_my_l\,dx = 0$ for $m \neq l$, and $\int ry_m^2\,dx = 1$. In the second sum on the right, the integral equals a_m, by (2) with $\|y_m\|^2 = 1$. Hence the first term on the right cancels half of the second term, so that the right side reduces to (analog of (6) in Sec. 11.4)

$$-\sum_{m=0}^{k} a_m^2 + \int_a^b rf^2\,dx.$$

This is nonnegative because in the previous formula the integrand on the left is nonnegative (recall that the weight $r(x)$ is positive!) and so is the integral on the left. This proves the important **Bessel's inequality** (analog of (7) in Sec. 11.4)

(16) $$\sum_{m=0}^{k} a_m^2 \leqq \|f\|^2 = \int_a^b r(x)f(x)^2\,dx \qquad (k = 1, 2, \cdots),$$

Here we can let $k \to \infty$, because the left sides form a monotone increasing sequence that is bounded by the right side, so that we have convergence by the familiar Theorem 1 in App. A.3.3 Hence

$$\sum_{m=0}^{\infty} a_m^2 \leqq \|f\|^2. \tag{17}$$

Furthermore, if $y_0, y_1, \cdots$ is complete in a set of functions S, then (13) holds for every f belonging to S. By (13) this implies equality in (16) with $k \to \infty$. Hence in the case of completeness every f in S saisfies the so-called **Parseval equality** (analog of (8) in Sec. 11.4)

$$\sum_{m=0}^{\infty} a_m^2 = \|f\|^2 = \int_a^b r(x) f(x)^2 \, dx. \tag{18}$$

As a consequence of (18) we prove that in the case of *completeness* there is no function orthogonal to *every* function of the orthonormal set, with the trivial exception of a function of zero norm:

THEOREM 2

Completeness

Let $y_0, y_1, \cdots$ be a complete orthonormal set on $a \leqq x \leqq b$ in a set of functions S. Then if a function f belongs to S and is orthogonal to every y_m, it must have norm zero. In particular, if f is continuous, then f must be identically zero.

PROOF Since f is orthogonal to every y_m, the left side of (18) must be zero. If f is continuous, then $\|f\| = 0$ implies $f(x) \equiv 0$, as can be seen directly from (5) in Sec. 11.5 with f instead of y_m because $r(x) > 0$ by assumption. ∎

PROBLEM SET 11.6

1–7 FOURIER–LEGENDRE SERIES

Showing the details, develop

1. $63x^5 - 90x^3 + 35x$
2. $(x + 1)^2$
3. $1 - x^4$
4. $1, \quad x, \quad x^2, \quad x^3, \quad x^4$
5. Prove that if $f(x)$ is even (is odd, respectively), its Fourier–Legendre series contains only $P_m(x)$ with even m (only $P_m(x)$ with odd m, respectively). Give examples.
6. What can you say about the coefficients of the Fourier–Legendre series of $f(x)$ if the Maclaurin series of $f(x)$ contains only powers x^{4m} $(m = 0, 1, 2, \cdots)$?
7. What happens to the Fourier–Legendre series of a *polynomial* $f(x)$ if you change a coefficient of $f(x)$? Experiment. Try to prove your answer.

8–13 CAS EXPERIMENT

FOURIER–LEGENDRE SERIES. Find and graph (on common axes) the partial sums up to S_{m_0} whose graph practically coincides with that of $f(x)$ within graphical accuracy. State m_0. On what does the size of m_0 seem to depend?

8. $f(x) = \sin \pi x$
9. $f(x) = \sin 2\pi x$
10. $f(x) = e^{-x^2}$
11. $f(x) = (1 + x^2)^{-1}$
12. $f(x) = J_0(\alpha_{0,1} x)$, $\alpha_{0,1}$ = the first positive zero of $J_0(x)$
13. $f(x) = J_0(\alpha_{0,2} x)$, $\alpha_{0,2}$ = the second positive zero of $J_0(x)$

14. TEAM PROJECT. Orthogonality on the Entire Real Axis. Hermite Polynomials.[8] These orthogonal polynomials are defined by $He_0(1) = 1$ and

$$(19) \quad He_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n}(e^{-x^2/2}), \qquad n = 1, 2, \cdots.$$

REMARK. As is true for many special functions, the literature contains more than one notation, and one sometimes defines as Hermite polynomials the functions

$$H_0^* = 1, \qquad H_n^*(x) = (-1)^n e^{x^2} \frac{d^n e^{-x^2}}{dx^n}.$$

This differs from our definition, which is preferred in applications.

(a) Small Values of *n*. Show that

$$He_1(x) = x, \qquad He_2(x) = x^2 - 1,$$
$$He_3(x) = x^3 - 3x, \qquad He_4(x) = x^4 - 6x^2 + 3.$$

(b) Generating Function. A generating function of the Hermite polynomials is

$$(20) \qquad e^{tx - t^2/2} = \sum_{n=0}^{\infty} a_n(x) t^n$$

because $He_n(x) = n!\, a_n(x)$. Prove this. *Hint*: Use the formula for the coefficients of a Maclaurin series and note that $tx - \frac{1}{2}t^2 = \frac{1}{2}x^2 - \frac{1}{2}(x - t)^2$.

(c) Derivative. Differentiating the generating function with respect to x, show that

$$(21) \qquad He_n'(x) = nHe_{n-1}(x).$$

(d) Orthogonality on the *x*-Axis needs a weight function that goes to zero sufficiently fast as $x \to \pm\infty$, (Why?) Show that the Hermite polynomials are orthogonal on $-\infty < x < \infty$ with respect to the weight function $r(x) = e^{-x^2/2}$. *Hint*. Use integration by parts and (21).

(e) ODEs. Show that

$$(22) \qquad He_n'(x) = xHe_n(x) - He_{n+1}(x).$$

Using this with $n - 1$ instead of n and (21), show that $y = He_n(x)$ satisfies the ODE

$$(23) \qquad y'' = xy' + ny = 0.$$

Show that $w = e^{-x^2/4}y$ is a solution of **Weber's equation**

$$(24) \quad w'' + (n + \tfrac{1}{2} - \tfrac{1}{4}x^2)w = 0 \qquad (n = 0, 1, \cdots).$$

15. CAS EXPERIMENT. Fourier–Bessel Series. Use Example 2 and $R = 1$, so that you get the series

$$(25) \qquad f(x) = a_1 J_0(\alpha_{0,1}x) + a_2 J_0(\alpha_{0,2}x) + a_3 J_0(\alpha_{0,3}x) + \cdots$$

With the zeros $\alpha_{0,1}\alpha_{0,2}, \cdots$ from your CAS (see also Table A1 in App. 5).

(a) Graph the terms $J_0(\alpha_{0,1}x), \cdots, J_0(\alpha_{0,10}x)$ for $0 \leqq x \leqq 1$ on common axes.

(b) Write a program for calculating partial sums of (25). Find out for what $f(x)$ your CAS can evaluate the integrals. Take two such $f(x)$ and comment empirically on the speed of convergence by observing the decrease of the coefficients.

(c) Take $f(x) = 1$ in (25) and evaluate the integrals for the coefficients analytically by (21a), Sec. 5.4, with $v = 1$. Graph the first few partial sums on common axes.

11.7 Fourier Integral

Fourier series are powerful tools for problems involving functions that are periodic or are of interest on a finite interval only. Sections 11.2 and 11.3 first illustrated this, and various further applications follow in Chap. 12. Since, of course, many problems involve functions that are ***nonperiodic*** *and are of interest on the whole x-axis,* we ask what can be done to extend the method of Fourier series to such functions. This idea will lead to "Fourier integrals."

In Example 1 we start from a special function f_L of period $2L$ and see what happens to its Fourier series if we let $L \to \infty$. Then we do the same for an *arbitrary* function f_L of period $2L$. This will motivate and suggest the main result of this section, which is an integral representation given in Theorem 1 below.

---

[8] CHARLES HERMITE (1822–1901), French mathematician, is known for his work in algebra and number theory. The great HENRI POINCARÉ (1854–1912) was one of his students.

EXAMPLE 1 Rectangular Wave

Consider the periodic rectangular wave $f_L(x)$ of period $2L > 2$ given by

$$f_L(x) = \begin{cases} 0 & \text{if} \quad -L < x < -1 \\ 1 & \text{if} \quad -1 < x < 1 \\ 0 & \text{if} \quad 1 < x < L. \end{cases}$$

The left part of Fig. 280 shows this function for $2L = 4, 8, 16$ as well as the nonperiodic function $f(x)$, which we obtain from f_L if we let $L \to \infty$,

$$f(x) = \lim_{L\to\infty} f_L(x) = \begin{cases} 1 & \text{if } -1 < x < 1 \\ 0 & \text{otherwise.} \end{cases}$$

We now explore what happens to the Fourier coefficients of f_L as L increases. Since f_L is even, $b_n = 0$ for all n. For a_n the Euler formulas (6), Sec. 11.2, give

$$a_0 = \frac{1}{2L}\int_{-1}^{1} dx = \frac{1}{L}, \qquad a_n = \frac{1}{L}\int_{-1}^{1} \cos\frac{n\pi x}{L}\, dx = \frac{2}{L}\int_{0}^{1} \cos\frac{n\pi x}{L}\, dx = \frac{2}{L}\,\frac{\sin(n\pi/L)}{n\pi/L}.$$

This sequence of Fourier coefficients is called the **amplitude spectrum** of f_L because $|a_n|$ is the maximum amplitude of the wave $a_n \cos(n\pi x/L)$. Figure 280 shows this spectrum for the periods $2L = 4, 8, 16$. We see that for increasing L these amplitudes become more and more dense on the positive w_n-axis, where $w_n = n\pi/L$. Indeed, for $2L = 4, 8, 16$ we have 1, 3, 7 amplitudes per "half-wave" of the function $(2 \sin w_n)/(Lw_n)$ (dashed in the figure). Hence for $2L = 2^k$ we have $2^{k-1} - 1$ amplitudes per half-wave, so that these amplitudes will eventually be everywhere dense on the positive w_n-axis (and will decrease to zero).

The outcome of this example gives an intuitive impression of what about to expect if we turn from our special function to an arbitrary one, as we shall do next. ■

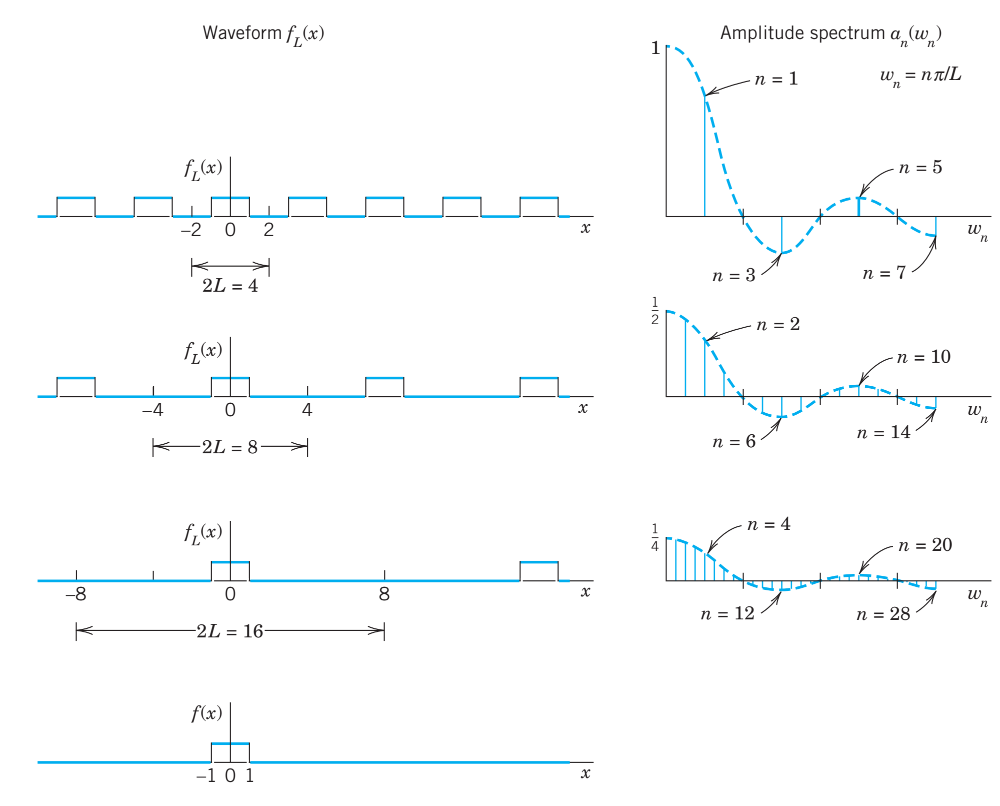


Fig. 280. Waveforms and amplitude spectra in Example 1

From Fourier Series to Fourier Integral

We now consider any periodic function $f_L(x)$ of period $2L$ that can be represented by a Fourier series

$$f_L(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos w_n x + b_n \sin w_n x), \qquad w_n = \frac{n\pi}{L}$$

and find out what happens if we let $L \to \infty$. Together with Example 1 the present calculation will suggest that we should expect an integral (instead of a series) involving $\cos wx$ and $\sin wx$ with w no longer restricted to integer multiples $w = w_n = n\pi/L$ of π/L but taking *all* values. We shall also see what form such an integral might have.

If we insert a_n and b_n from the Euler formulas (6), Sec. 11.2, and denote the variable of integration by v, the Fourier series of $f_L(x)$ becomes

$$f_L(x) = \frac{1}{2L}\int_{-L}^{L} f_L(v)\, dv + \frac{1}{L}\sum_{n=1}^{\infty}\left[\cos w_n x \int_{-L}^{L} f_L(v)\cos w_n v\, dv + \sin w_n x \int_{-L}^{L} f_L(v)\sin w_n v\, dv\right].$$

We now set

$$\Delta w = w_{n+1} - w_n = \frac{(n+1)\pi}{L} - \frac{n\pi}{L} = \frac{\pi}{L}.$$

Then $1/L = \Delta w/\pi$, and we may write the Fourier series in the form

$$f_L(x) = \frac{1}{2L}\int_{-L}^{L} f_L(v)\, dv + \frac{1}{\pi}\sum_{n=1}^{\infty}\left[(\cos w_n x)\,\Delta w \int_{-L}^{L} f_L(v)\cos w_n v\, dv + (\sin w_n x)\Delta w \int_{-L}^{L} f_L(v)\sin w_n v\, dv\right]. \tag{1}$$

This representation is valid for any fixed L, arbitrarily large, but finite.

We now let $L \to \infty$ and assume that the resulting nonperiodic function

$$f(x) = \lim_{L\to\infty} f_L(x)$$

is **absolutely integrable** on the x-axis; that is, the following (finite!) limits exist:

$$\lim_{a\to-\infty}\int_{a}^{0} |f(x)|\, dx + \lim_{b\to\infty}\int_{0}^{b} |f(x)|\, dx \quad \left(\text{written} \int_{-\infty}^{\infty} |f(x)|\, dx\right). \tag{2}$$

Then $1/L \to 0$, and the value of the first term on the right side of (1) approaches zero. Also $\Delta w = \pi/L \to 0$ and it seems ***plausible*** that the infinite series in (1) becomes an

integral from 0 to ∞, which represents $f(x)$, namely,

$$(3) \qquad f(x) = \frac{1}{\pi}\int_0^{\infty}\left[\cos wx \int_{-\infty}^{\infty} f(v)\cos wv\, dv + \sin wx \int_{-\infty}^{\infty} f(v)\sin wv\, dv\right] dw.$$

If we introduce the notations

$$\textbf{(4)} \qquad A(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos wv\, dv, \qquad B(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\sin wv\, dv$$

we can write this in the form

$$\textbf{(5)} \qquad f(x) = \int_0^{\infty} [A(w)\cos wx + B(w)\sin wx]\, dw.$$

This is called a representation of $f(x)$ by a **Fourier integral**.

It is clear that our naive approach merely ***suggests*** the representation (5), but by no means establishes it; in fact, the limit of the series in (1) as Δw approaches zero is not the definition of the integral (3). Sufficient conditions for the validity of (5) are as follows.

THEOREM 1

Fourier Integral

If $f(x)$ is piecewise continuous (see Sec. 6.1) *in every finite interval and has a right-hand derivative and a left-hand derivative at every point* (see Sec 11.1) *and if the integral* (2) *exists, then $f(x)$ can be represented by a Fourier integral* (5) *with A and B given by* (4)*. At a point where $f(x)$ is discontinuous the value of the Fourier integral equals the average of the left- and right-hand limits of $f(x)$ at that point* (see Sec. 11.1). (Proof in Ref. [C12]; see App. 1.)

Applications of Fourier Integrals

The main application of Fourier integrals is in solving ODEs and PDEs, as we shall see for PDEs in Sec. 12.6. However, we can also use Fourier integrals in integration and in discussing functions defined by integrals, as the next example.

EXAMPLE 2 Single Pulse, Sine Integral. Dirichlet's Discontinuous Factor. Gibbs Phenomenon

Find the Fourier integral representation of the function

$$f(x) = \begin{cases} 1 & \text{if } |x| < 1 \\ 0 & \text{if } |x| > 1 \end{cases} \qquad \text{(Fig. 281)}$$

f(x)
1
–1 0 1 x

Fig. 281. Example 2

Solution. From (4) we obtain

$$A(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos wv \, dv = \frac{1}{\pi}\int_{-1}^{1} \cos wv \, dv = \frac{\sin wv}{\pi w}\bigg|_{-1}^{1} = \frac{2\sin w}{\pi w}$$

$$B(w) = \frac{1}{\pi}\int_{-1}^{1} \sin wv \, dv = 0$$

and (5) gives the *answer*

$$f(x) = \frac{2}{\pi}\int_{0}^{\infty} \frac{\cos wx \sin w}{w} \, dw. \tag{6}$$

The average of the left- and right-hand limits of $f(x)$ at $x = 1$ is equal to $(1 + 0)/2$, that is, $\frac{1}{2}$.

Furthermore, from (6) and Theorem 1 we obtain (multiply by $\pi/2$)

$$\int_{0}^{\infty} \frac{\cos wx \sin w}{w} \, dw = \begin{cases} \pi/2 & \text{if} \quad 0 \leqq x < 1, \\ \pi/4 & \text{if} \quad x = 1, \\ 0 & \text{if} \quad x > 1. \end{cases} \tag{7}$$

We mention that this integral is called **Dirichlet's discontinous factor**. (For P. L. Dirichlet see Sec. 10.8.)

The case $x = 0$ is of particular interest. If $x = 0$, then (7) gives

$$\int_{0}^{\infty} \frac{\sin w}{w} \, dw = \frac{\pi}{2}. \tag{8*}$$

We see that this integral is the limit of the so-called **sine integral**

$$\text{Si}(u) = \int_{0}^{u} \frac{\sin w}{w} \, dw \tag{8}$$

as $u \to \infty$. The graphs of Si(u) and of the integrand are shown in Fig. 282.

In the case of a Fourier series the graphs of the partial sums are approximation curves of the curve of the periodic function represented by the series. Similarly, in the case of the Fourier integral (5), approximations are obtained by replacing ∞ by numbers a. Hence the integral

$$\frac{2}{\pi}\int_{0}^{a} \frac{\cos wx \sin w}{w} \, dw \tag{9}$$

approximates the right side in (6) and therefore $f(x)$.

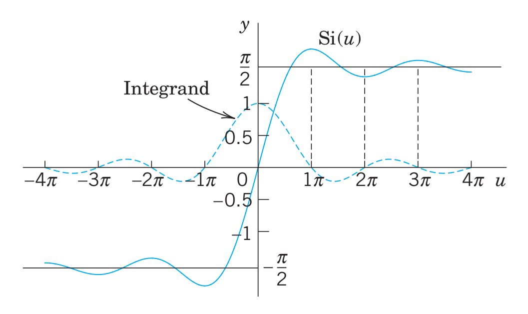


Fig. 282. Sine integral Si(u) and integrand

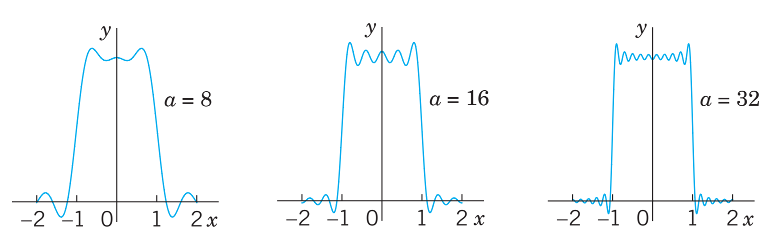


Fig. 283. The integral (9) for $a = 8, 16$, and 32, illustrating the development of the Gibbs phenomenon

Figure 283 shows oscillations near the points of discontinuity of $f(x)$. We might expect that these oscillations disappear as a approaches infinity. But this is not true; with increasing a, they are shifted closer to the points $x = \pm 1$. This unexpected behavior, which also occurs in connection with Fourier series (see Sec. 11.2), is known as the **Gibbs phenomenon**. We can explain it by representing (9) in terms of sine integrals as follows. Using (11) in App. A3.1, we have

$$\frac{2}{\pi}\int_0^a \frac{\cos wx \sin w}{w}\, dw = \frac{1}{\pi}\int_0^a \frac{\sin (w + wx)}{w}\, dw + \frac{1}{\pi}\int_0^a \frac{\sin (w - wx)}{w}\, dw.$$

In the first integral on the right we set $w + wx = t$. Then $dw/w = dt/t$, and $0 \leqq w \leqq a$ corresponds to $0 \leqq t \leqq (x + 1)a$. In the last integral we set $w - wx = -t$. Then $dw/w = dt/t$, and $0 \leqq w \leqq a$ corresponds to $0 \leqq t \leqq (x - 1)a$. Since $\sin(-t) = -\sin t$, we thus obtain

$$\frac{2}{\pi}\int_0^a \frac{\cos wx \sin w}{w}\, dw = \frac{1}{\pi}\int_0^{(x+1)a} \frac{\sin t}{t}\, dt - \frac{1}{\pi}\int_0^{(x-1)a} \frac{\sin t}{t}\, dt.$$

From this and (8) we see that our integral (9) equals

$$\frac{1}{\pi}\,\text{Si}(a[x + 1]) - \frac{1}{\pi}\,\text{Si}(a[x - 1])$$

and the oscillations in Fig. 283 result from those in Fig. 282. The increase of a amounts to a transformation of the scale on the axis and causes the shift of the oscillations (the waves) toward the points of discontinuity -1 and 1. ■

Fourier Cosine Integral and Fourier Sine Integral

Just as Fourier *series* simplify if a function is even or odd (see Sec. 11.2), so do Fourier *integrals*, and you can save work. Indeed, if f has a Fourier integral representation and is *even*, then $B(w) = 0$ in (4). This holds because the integrand of $B(w)$ is odd. Then (5) reduces to a **Fourier cosine integral**

(10) $$f(x) = \int_0^\infty A(w) \cos wx \, dw \qquad \text{where} \qquad A(w) = \frac{2}{\pi}\int_0^\infty f(v) \cos wv \, dv.$$

Note the change in $A(w)$: for even f the integrand is even, hence the integral from $-\infty$ to ∞ equals twice the integral from 0 to ∞, just as in (7a) of Sec. 11.2.

Similarly, if f has a Fourier integral representation and is *odd*, then $A(w) = 0$ in (4). This is true because the integrand of $A(w)$ is odd. Then (5) becomes a **Fourier sine integral**

(11) $$f(x) = \int_0^\infty B(w) \sin wx \, dw \qquad \text{where} \qquad B(w) = \frac{2}{\pi}\int_0^\infty f(v) \sin wv \, dv.$$

Note the change of $B(w)$ to an integral from 0 to ∞ because $B(w)$ is even (odd times odd is even).

Earlier in this section we pointed out that the main application of the Fourier integral representation is in differential equations. However, these representations also help in evaluating integrals, as the following example shows for integrals from 0 to ∞.

EXAMPLE 3 Laplace Integrals

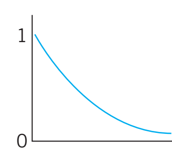


Fig. 284. $f(x)$ in Example 3

We shall derive the Fourier cosine and Fourier sine integrals of $f(x) = e^{-kx}$, where $x > 0$ and $k > 0$ (Fig. 284). The result will be used to evaluate the so-called Laplace integrals.

Solution. **(a)** From (10) we have $A(w) = \frac{2}{\pi}\int_0^\infty e^{-kv}\cos wv\, dv$. Now, by integration by parts,

$$\int e^{-kv}\cos wv\, dv = -\frac{k}{k^2+w^2}e^{-kv}\left(-\frac{w}{k}\sin wv + \cos wv\right).$$

If $v = 0$, the expression on the right equals $-k/(k^2 + w^2)$. If v approaches infinity, that expression approaches zero because of the exponential factor. Thus $2/\pi$ times the integral from 0 to ∞ gives

$$A(w) = \frac{2k/\pi}{k^2+w^2}. \tag{12}$$

By substituting this into the first integral in (10) we thus obtain the Fourier cosine integral representation

$$f(x) = e^{-kx} = \frac{2k}{\pi}\int_0^\infty \frac{\cos wx}{k^2+w^2}\, dw \qquad (x > 0, \quad k > 0).$$

From this representation we see that

$$\int_0^\infty \frac{\cos wx}{k^2+w^2}\, dw = \frac{\pi}{2k}e^{-kx} \qquad (x > 0, \quad k > 0). \tag{13}$$

(b) Similarly, from (11) we have $B(w) = \frac{2}{\pi}\int_0^\infty e^{-kv}\sin wv\, dv$. By integration by parts,

$$\int e^{-kv}\sin wv\, dv = -\frac{w}{k^2+w^2}e^{-kv}\left(\frac{k}{w}\sin wv + \cos wv\right).$$

This equals $-w/(k^2 + w^2)$ if $v = 0$, and approaches 0 as $v \to \infty$. Thus

$$B(w) = \frac{2w/\pi}{k^2+w^2}. \tag{14}$$

From (14) we thus obtain the Fourier sine integral representation

$$f(x) = e^{-kx} = \frac{2}{\pi}\int_0^\infty \frac{w \sin wx}{k^2+w^2}\, dw.$$

From this we see that

$$\int_0^\infty \frac{w\sin wx}{k^2+w^2}\, dw = \frac{\pi}{2}e^{-kx} \qquad (x > 0, \quad k > 0). \tag{15}$$

The integrals (13) and (15) are called the **Laplace integrals**. ■

PROBLEM SET 11.7

1–6 EVALUATION OF INTEGRALS

Show that the integral represents the indicated function. *Hint*. Use (5), (10), or (11); the integral tells you which one, and its value tells you what function to consider. Show your work in detail.

1. $$\int_0^\infty \frac{\cos xw + w \sin xw}{1 + w^2}\, dx = \begin{cases} 0 & \text{if } x < 0 \\ \pi/2 & \text{if } x = 0 \\ \pi e^{-x} & \text{if } x > 0 \end{cases}$$

2. $$\int_0^\infty \frac{\sin \pi w \sin xw}{1 - w^2}\, dw = \begin{cases} \frac{\pi}{2} \sin x & \text{if } 0 \leqq x \leqq \pi \\ 0 & \text{if } x > \pi \end{cases}$$

3. $$\int_0^\infty \frac{1 - \cos \pi w}{w} \sin xw\, dw = \begin{cases} \frac{1}{2}\pi & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$$

4. $$\int_0^\infty \frac{\cos \frac{1}{2}\pi w}{1 - w^2} \cos xw\, dw = \begin{cases} \frac{1}{2}\pi \cos x & \text{if } 0 < |x| < \frac{1}{2}\pi \\ 0 & \text{if } |x| \geqq \frac{1}{2}\pi \end{cases}$$

5. $$\int_0^\infty \frac{\sin w - w \cos w}{w^2} \sin xw\, dw = \begin{cases} \frac{1}{2}\pi x & \text{if } 0 < x < 1 \\ \frac{1}{4}\pi & \text{if } x = 1 \\ 0 & \text{if } x > 1 \end{cases}$$

6. $$\int_0^\infty \frac{w^3 \sin xw}{w^4 + 4}\, dw = \tfrac{1}{2}\pi e^{-x} \cos x \quad \text{if } x > 0$$

7–12 FOURIER COSINE INTEGRAL REPRESENTATIONS

Represent $f(x)$ as an integral (10).

7. $f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

8. $f(x) = \begin{cases} x^2 & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

9. $f(x) = 1/(1 + x^2)$ [$x > 0$. *Hint*. See (13).]

10. $f(x) = \begin{cases} a^2 - x^2 & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

11. $f(x) = \begin{cases} \sin x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

12. $f(x) = \begin{cases} e^{-x} & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

13. **CAS EXPERIMENT. Approximate Fourier Cosine Integrals.** Graph the integrals in Prob. 7, 9, and 11 as functions of x. Graph approximations obtained by replacing ∞ with finite upper limits of your choice. Compare the quality of the approximations. Write a short report on your empirical results and observations.

14. **PROJECT. Properties of Fourier Integrals**

(a) Fourier cosine integral. Show that (10) implies

(a1) $$f(ax) = \frac{1}{a}\int_0^\infty A\left(\frac{w}{a}\right) \cos xw\, dw$$
$(a > 0)$ (*Scale change*)

(a2) $$xf(x) = \int_0^\infty B^*(w) \sin xw\, dw,$$
$$B^* = -\frac{dA}{dw}, \qquad A \text{ as in (10)}$$

(a3) $$x^2 f(x) = \int_0^\infty A^*(w) \cos xw\, dw,$$
$$A^* = -\frac{d^2A}{dw^2}.$$

(b) Solve Prob. 8 by applying (a3) to the result of Prob. 7.

(c) Verify (a2) for $f(x) = 1$ if $0 < x < a$ and $f(x) = 0$ if $x > a$.

(d) Fourier sine integral. Find formulas for the Fourier sine integral similar to those in (a).

15. **CAS EXPERIMENT. Sine Integral.** Plot Si(u) for positive u. Does the sequence of the maximum and minimum values give the impression that it converges and has the limit $\pi/2$? Investigate the Gibbs phenomenon graphically.

16–20 FOURIER SINE INTEGRAL REPRESENTATIONS

Represent $f(x)$ as an integral (11).

16. $f(x) = \begin{cases} x & \text{if } 0 < x < a \\ 0 & \text{if } x > a \end{cases}$

17. $f(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

18. $f(x) = \begin{cases} \cos x & \text{if } 0 < x < \pi \\ 0 & \text{if } x > \pi \end{cases}$

19. $f(x) = \begin{cases} e^x & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

20. $f(x) = \begin{cases} e^{-x} & \text{if } 0 < x < 1 \\ 0 & \text{if } x > 1 \end{cases}$

11.8 Fourier Cosine and Sine Transforms

An **integral transform** is a transformation in the form of an integral that produces from given functions new functions depending on a different variable. One is mainly interested in these transforms because they can be used as tools in solving ODEs, PDEs, and integral equations and can often be of help in handling and applying special functions. The Laplace transform of Chap. 6 serves as an example and is by far the most important integral transform in engineering.

Next in order of importance are Fourier transforms. They can be obtained from the Fourier integral in Sec. 11.7 in a straightforward way. In this section we derive two such transforms that are real, and in Sec. 11.9 a complex one.

Fourier Cosine Transform

The Fourier cosine transform concerns ***even functions*** $f(x)$. We obtain it from the Fourier cosine integral [(10) in Sec. 10.7]

$$f(x) = \int_0^\infty A(w) \cos wx \, dw, \qquad \text{where} \qquad A(w) = \frac{2}{\pi} \int_0^\infty f(v) \cos wv \, dv.$$

Namely, we set $A(w) = \sqrt{2/\pi}\, \hat{f}_c(w)$, where c suggests "cosine." Then, writing $v = x$ in the formula for $A(w)$, we have

(1a) $$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \cos wx \, dx$$

and

(1b) $$f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_c(w) \cos wx \, dw.$$

Formula (1a) gives from $f(x)$ a new function $\hat{f}_c(w)$, called the **Fourier cosine transform** of $f(x)$. Formula (1b) gives us back $f(x)$ from $\hat{f}_c(w)$, and we therefore call $f(x)$ the **inverse Fourier cosine transform** of $\hat{f}_c(w)$.

The process of obtaining the transform $\hat{f}_c$ from a given f is also called the **Fourier cosine transform** or the *Fourier cosine transform method.*

Fourier Sine Transform

Similarly, in (11), Sec. 11.7, we set $B(w) = \sqrt{2/\pi}\, \hat{f}_s(w)$, where s suggests "sine." Then, writing $v = x$, we have from (11), Sec. 11.7, the **Fourier sine transform**, of $f(x)$ given by

(2a) $$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^\infty f(x) \sin wx \, dx,$$

and the **inverse Fourier sine transform** of $\hat{f}_s(w)$, given by

(2b)
$$f(x) = \sqrt{\frac{2}{\pi}} \int_0^\infty \hat{f}_s(w) \sin wx \, dw.$$

The process of obtaining $f_s(w)$ from $f(x)$ is also called the **Fourier sine transform** or the *Fourier sine transform method.*

Other notations are

$$\mathscr{F}_c(f) = \hat{f}_c, \qquad \mathscr{F}_s(f) = \hat{f}_s$$

and $\mathscr{F}_c^{-1}$ and $\mathscr{F}_s^{-1}$ for the inverses of $\mathscr{F}_c$ and $\mathscr{F}_s$, respectively.

EXAMPLE 1 **Fourier Cosine and Fourier Sine Transforms**

Find the Fourier cosine and Fourier sine transforms of the function

$$f(x) = \begin{cases} k & \text{if} \quad 0 < x < a \\ 0 & \text{if} \quad x > a \end{cases} \qquad \text{(Fig. 285).}$$

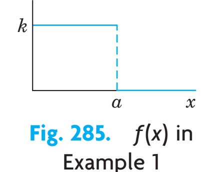


Fig. 285. $f(x)$ in Example 1

Solution. From the definitions (1a) and (2a) we obtain by integration

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}}\, k \int_0^a \cos wx \, dx = \sqrt{\frac{2}{\pi}}\, k \left(\frac{\sin aw}{w}\right)$$

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}}\, k \int_0^a \sin wx \, dx = \sqrt{\frac{2}{\pi}}\, k \left(\frac{1 - \cos aw}{w}\right).$$

This agrees with formulas 1 in the first two tables in Sec. 11.10 (where $k = 1$).

Note that for $f(x) = k = \text{const}$ $(0 < x < \infty)$, these transforms do not exist. (Why?) ■

EXAMPLE 2 **Fourier Cosine Transform of the Exponential Function**

Find $\mathscr{F}_c(e^{-x})$.

Solution. By integration by parts and recursion,

$$\mathscr{F}_c(e^{-x}) = \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-x} \cos wx \, dx = \sqrt{\frac{2}{\pi}} \frac{e^{-x}}{1 + w^2}(-\cos wx + w \sin wx)\bigg|_0^\infty = \frac{\sqrt{2/\pi}}{1 + w^2}.$$

This agrees with formula 3 in Table I, Sec. 11.10, with $a = 1$. See also the next example. ■

What did we do to introduce the two integral transforms under consideration? Actually not much: We changed the notations A and B to get a "symmetric" distribution of the constant $2/\pi$ in the original formulas (1) and (2). This redistribution is a standard convenience, but it is not essential. One could do without it.

What have we gained? We show next that these transforms have operational properties that permit them to convert differentiations into algebraic operations (just as the Laplace transform does). This is the key to their application in solving differential equations.

Linearity, Transforms of Derivatives

If $f(x)$ is absolutely integrable (see Sec. 11.7) on the positive x-axis and piecewise continuous (see Sec. 6.1) on every finite interval, then the Fourier cosine and sine transforms of f exist.

Furthermore, if f and g have Fourier cosine and sine transforms, so does $af + bg$ for any constants a and b, and by (1a)

$$\begin{aligned}\mathcal{F}_c(af + bg) &= \sqrt{\frac{2}{\pi}}\int_0^\infty [af(x) + bg(x)] \cos wx\, dx \\ &= a\sqrt{\frac{2}{\pi}}\int_0^\infty f(x) \cos wx\, dx + b\sqrt{\frac{2}{\pi}}\int_0^\infty g(x) \cos wx\, dx.\end{aligned}$$

The right side is $a\mathcal{F}_c(f) + b\mathcal{F}_c(g)$. Similarly for $\mathcal{F}_s$, by (2). This shows that the Fourier cosine and sine transforms are **linear operations**,

$$\text{(3)} \qquad \begin{aligned}&\text{(a)} \quad \mathcal{F}_c(af + bg) = a\mathcal{F}_c(f) + b\mathcal{F}_c(g),\\ &\text{(b)} \quad \mathcal{F}_s(af + bg) = a\mathcal{F}_s(f) + b\mathcal{F}_s(g).\end{aligned}$$

THEOREM 1

Cosine and Sine Transforms of Derivatives

Let $f(x)$ be continuous and absolutely integrable on the x-axis, let $f'(x)$ be piecewise continuous on every finite interval, and let $f(x) \to 0$ as $x \to \infty$. Then

$$\textbf{(4)} \qquad \begin{aligned}&\textbf{(a)} \quad \mathcal{F}_c\{f'(x)\} = w\mathcal{F}_s\{f(x)\} - \sqrt{\frac{2}{\pi}}f(0),\\ &\textbf{(b)} \quad \mathcal{F}_s\{f'(x)\} = -w\mathcal{F}_c\{f(x)\}.\end{aligned}$$

PROOF This follows from the definitions and by using integration by parts, namely,

$$\begin{aligned}\mathcal{F}_c\{f'(x)\} &= \sqrt{\frac{2}{\pi}}\int_0^\infty f'(x) \cos wx\, dx \\ &= \sqrt{\frac{2}{\pi}}\left[f(x)\cos wx\Big|_0^\infty + w\int_0^\infty f(x) \sin wx\, dx\right] \\ &= -\sqrt{\frac{2}{\pi}}f(0) + w\mathcal{F}_s\{f(x)\};\end{aligned}$$

and similarly,

$$\begin{aligned}\mathcal{F}_s\{f'(x)\} &= \sqrt{\frac{2}{\pi}}\int_0^\infty f'(x) \sin wx\, dx \\ &= \sqrt{\frac{2}{\pi}}\left[f(x)\sin wx\Big|_0^\infty - w\int_0^\infty f(x) \cos wx\, dx\right] \\ &= 0 - w\mathcal{F}_c\{f(x)\}.\end{aligned}$$

■

Formula (4a) with f' instead of f gives (when f', f'' satisfy the respective assumptions for f, f' in Theorem 1)

$$\mathcal{F}_c\{f''(x)\} = w\mathcal{F}_s\{f'(x)\} - \sqrt{\frac{2}{\pi}}f'(0);$$

hence by (4b)

(5a) $$\mathcal{F}_c\{f''(x)\} = -w^2\mathcal{F}_c\{f(x)\} - \sqrt{\frac{2}{\pi}}f'(0).$$

Similarly,

(5b) $$\mathcal{F}_s\{f''(x)\} = -w^2\mathcal{F}_s\{f(x)\} + \sqrt{\frac{2}{\pi}}\,wf(0).$$

A basic application of (5) to PDEs will be given in Sec. 12.7. For the time being we show how (5) can be used for deriving transforms.

EXAMPLE 3 An Application of the Operational Formula (5)

Find the Fourier cosine transform $\mathcal{F}_c(e^{-ax})$ of $f(x) = e^{-ax}$, where $a > 0$.

Solution. By differentiation, $(e^{-ax})'' = a^2e^{-ax}$; thus

$$a^2f(x) = f''(x).$$

From this, (5a), and the linearity (3a),

$$\begin{aligned} a^2\mathcal{F}_c(f) &= \mathcal{F}_c(f'') \\ &= -w^2\mathcal{F}_c(f) - \sqrt{\frac{2}{\pi}}f'(0) \\ &= -w^2\mathcal{F}_c(f) + a\sqrt{\frac{2}{\pi}}. \end{aligned}$$

Hence

$$(a^2 + w^2)\mathcal{F}_c(f) = a\sqrt{2/\pi}.$$

The *answer* is (see Table I, Sec. 11.10)

$$\mathcal{F}_c(e^{-ax}) = \sqrt{\frac{2}{\pi}}\left(\frac{a}{a^2 + w^2}\right) \qquad (a > 0).$$ ■

Tables of Fourier cosine and sine transforms are included in Sec. 11.10.

PROBLEM SET 11.8

1–8 FOURIER COSINE TRANSFORM

1. Find the cosine transform $\hat{f}_c(w)$ of $f(x) = 1$ if $0 < x < 1$, $f(x) = -1$ if $1 < x < 2$, $f(x) = 0$ if $x > 2$.

2. Find f in Prob. 1 from the answer $\hat{f}_c$.

3. Find $\hat{f}_c(w)$ for $f(x) = x$ if $0 < x < 2$, $f(x) = 0$ if $x > 2$.

4. Derive formula 3 in Table I of Sec. 11.10 by integration.

5. Find $\hat{f}_c(w)$ for $f(x) = x^2$ if $0 < x < 1$, $f(x) = 0$ if $x > 1$.

6. Continuity assumptions. Find $\hat{g}_c(w)$ for $g(x) = 2$ if $0 < x < 1$, $g(x) = 0$ if $x > 1$. Try to obtain from it $\hat{f}_c(w)$ for $f(x)$ in Prob. 5 by using (5a).

7. Existence? Does the Fourier cosine transform of $x^{-1} \sin x$ $(0 < x < \infty)$ exist? Of $x^{-1} \cos x$? Give reasons.

8. Existence? Does the Fourier cosine transform of $f(x) = k = \text{const}$ $(0 < x < \infty)$ exist? The Fourier sine transform?

9–15 FOURIER SINE TRANSFORM

9. Find $\mathscr{F}_s(e^{-ax})$, $a > 0$, by integration.

10. Obtain the answer to Prob. 9 from (5b).

11. Find $f_s(w)$ for $f(x) = x^2$ if $0 < x < 1$, $f(x) = 0$ if $x > 1$.

12. Find $\mathscr{F}_s(xe^{-x^2/2})$ from (4b) and a suitable formula in Table I of Sec. 11.10.

13. Find $\mathscr{F}_s(e^{-x})$ from (4a) and formula 3 of Table I in Sec. 11.10.

14. Gamma function. Using formulas 2 and 4 in Table II of Sec. 11.10, prove $\Gamma(\tfrac{1}{2}) = \sqrt{\pi}$ [(30) in App. A3.1], a value needed for Bessel functions and other applications.

15. WRITING PROJECT. Finding Fourier Cosine and Sine Transforms. Write a short report on ways of obtaining these transforms, with illustrations by examples of your own.

11.9 Fourier Transform. Discrete and Fast Fourier Transforms

In Sec. 11.8 we derived two real transforms. Now we want to derive a complex transform that is called the **Fourier transform**. It will be obtained from the complex Fourier integral, which will be discussed next.

Complex Form of the Fourier Integral

The (real) Fourier integral is [see (4), (5), Sec. 11.7]

$$f(x) = \int_0^\infty [A(w) \cos wx + B(w) \sin wx]\, dw$$

where

$$A(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \cos wv\, dv, \qquad B(w) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \sin wv\, dv.$$

Substituting A and B into the integral for f, we have

$$f(x) = \frac{1}{\pi} \int_0^\infty \int_{-\infty}^{\infty} f(v)[\cos wv \cos wx + \sin wv \sin wx]\, dv\, dw.$$

By the addition formula for the cosine [(6) in App. A3.1] the expression in the brackets $[\cdots]$ equals $\cos(wv - wx)$ or, since the cosine is even, $\cos(wx - wv)$. We thus obtain

$$f(x) = \frac{1}{\pi}\int_0^{\infty}\left[\int_{-\infty}^{\infty} f(v)\cos(wx - wv)dv\right]dw. \tag{1*}$$

The integral in brackets is an *even* function of w, call it $F(w)$, because $\cos(wx - wv)$ is an even function of w, the function f does not depend on w, and we integrate with respect to v (not w). Hence the integral of $F(w)$ from $w = 0$ to ∞ is $\frac{1}{2}$ times the integral of $F(w)$ from $-\infty$ to ∞. Thus (note the change of the integration limit!)

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(v)\cos(wx - wv)\,dv\right]dw. \tag{1}$$

We claim that the integral of the form (1) with sin instead of cos is zero:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} f(v)\sin(wx - wv)\,dv\right]dw = 0. \tag{2}$$

This is true since $\sin(wx - wv)$ is an odd function of w, which makes the integral in brackets an odd function of w, call it $G(w)$. Hence the integral of $G(w)$ from $-\infty$ to ∞ is zero, as claimed.

We now take the integrand of (1) plus $i\,(= \sqrt{-1})$ times the integrand of (2) and use the **Euler formula** [(11) in Sec. 2.2]

$$e^{ix} = \cos x + i\sin x. \tag{3}$$

Taking $wx - wv$ instead of x in (3) and multiplying by $f(v)$ gives

$$f(v)\cos(wx - wv) + if(v)\sin(wx - wv) = f(v)e^{i(wx-wv)}.$$

Hence the result of adding (1) plus i times (2), called the **complex Fourier integral**, is

$$f(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(v)e^{iw(x-v)}\,dv\,dw \qquad (i = \sqrt{-1}). \tag{4}$$

To obtain the desired Fourier transform will take only a very short step from here.

Fourier Transform and Its Inverse

Writing the exponential function in (4) as a product of exponential functions, we have

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\left[\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(v)e^{-iwv}\,dv\right]e^{iwx}\,dw. \tag{5}$$

The expression in brackets is a function of w, is denoted by $\hat{f}(w)$, and is called the **Fourier transform** of f; writing $v = x$, we have

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-iwx}\,dx. \tag{6}$$

With this, (5) becomes

(7)
$$f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(w) e^{iwx}\, dw$$

and is called the **inverse Fourier transform** of $\hat{f}(w)$.

Another notation for the Fourier transform is

$$\hat{f} = \mathcal{F}(f),$$

so that

$$f = \mathcal{F}^{-1}(\hat{f}).$$

The process of obtaining the Fourier transform $\mathcal{F}(f) = \hat{f}$ from a given f is also called the **Fourier transform** or the *Fourier transform method*.

Using concepts defined in Secs. 6.1 and 11.7 we now state (without proof) conditions that are sufficient for the existence of the Fourier transform.

THEOREM 1

Existence of the Fourier Transform

If $f(x)$ is absolutely integrable on the x-axis and piecewise continuous on every finite interval, then the Fourier transform $\hat{f}(w)$ of $f(x)$ given by (6) *exists.*

EXAMPLE 1

Fourier Transform

Find the Fourier transform of $f(x) = 1$ if $|x| < 1$ and $f(x) = 0$ otherwise.

Solution. Using (6) and integrating, we obtain

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}} \int_{-1}^{1} e^{-iwx}\, dx = \frac{1}{\sqrt{2\pi}} \cdot \frac{e^{-iwx}}{-iw}\bigg|_{-1}^{1} = \frac{1}{-iw\sqrt{2\pi}}(e^{-iw} - e^{iw}).$$

As in (3) we have $e^{iw} = \cos w + i \sin w$, $e^{-iw} = \cos w - i \sin w$, and by subtraction

$$e^{iw} - e^{-iw} = 2i \sin w.$$

Substituting this in the previous formula on the right, we see that i drops out and we obtain the answer

$$\hat{f}(w) = \sqrt{\frac{\pi}{2}}\, \frac{\sin w}{w}.$$ ■

EXAMPLE 2

Fourier Transform

Find the Fourier transform $\mathcal{F}(e^{-ax})$ of $f(x) = e^{-ax}$ if $x > 0$ and $f(x) = 0$ if $x < 0$; here $a > 0$.

Solution. From the definition (6) we obtain by integration

$$\mathcal{F}(e^{-ax}) = \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{-ax} e^{-iwx}\, dx$$

$$= \frac{1}{\sqrt{2\pi}} \frac{e^{-(a+iw)x}}{-(a + iw)}\bigg|_{x=0}^{\infty} = \frac{1}{\sqrt{2\pi}(a + iw)}.$$

This proves formula 5 of Table III in Sec. 11.10. ■

Physical Interpretation: Spectrum

The nature of the representation (7) of $f(x)$ becomes clear if we think of it as a superposition of sinusoidal oscillations of all possible frequencies, called a **spectral representation**. This name is suggested by optics, where light is such a superposition of colors (frequencies). In (7), the "**spectral density**" $\hat{f}(w)$ measures the intensity of $f(x)$ in the frequency interval between w and $w + \Delta w$ (Δw small, fixed). We claim that, in connection with vibrations, the integral

$$\int_{-\infty}^{\infty} |\hat{f}(w)|^2 \, dw$$

can be interpreted as the **total energy** of the physical system. Hence an integral of $|\hat{f}(w)|^2$ from a to b gives the contribution of the frequencies w between a and b to the total energy.

To make this plausible, we begin with a mechanical system giving a single frequency, namely, the harmonic oscillator (mass on a spring, Sec. 2.4)

$$my'' + ky = 0.$$

Here we denote time t by x. Multiplication by y' gives $my'y'' + ky'y = 0$. By integration,

$$\tfrac{1}{2}mv^2 + \tfrac{1}{2}ky^2 = E_0 = \text{const}$$

where $v = y'$ is the velocity. The first term is the kinetic energy, the second the potential energy, and E_0 the total energy of the system. Now a general solution is (use (3) in Sec. 11.4 with $t = x$)

$$y = a_1 \cos w_0 x + b_1 \sin w_0 x = c_1 e^{iw_0 x} + c_{-1} e^{-iw_0 x}, \qquad w_0^2 = k/m$$

where $c_1 = (a_1 - ib_1)/2$, $c_{-1} = \bar{c}_1 = (a_1 + ib_1)/2$. We write simply $A = c_1 e^{iw_0 x}$, $B = c_{-1} e^{-iw_0 x}$. Then $y = A + B$. By differentiation, $v = y' = A' + B' = iw_0(A - B)$. Substitution of v and y on the left side of the equation for E_0 gives

$$E_0 = \tfrac{1}{2}mv^2 + \tfrac{1}{2}ky^2 = \tfrac{1}{2}m(iw_0)^2(A - B)^2 + \tfrac{1}{2}k(A + B)^2.$$

Here $w_0^2 = k/m$, as just stated; hence $mw_0^2 = k$. Also $i^2 = -1$, so that

$$E_0 = \tfrac{1}{2}k[-(A - B)^2 + (A + B)^2] = 2kAB = 2kc_1 e^{iw_0 x} c_{-1} e^{-iw_0 x} = 2kc_1 c_{-1} = 2k|c_1|^2.$$

Hence *the energy is proportional to the square of the amplitude* $|c_1|$.

As the next step, if a more complicated system leads to a periodic solution $y = f(x)$ that can be represented by a Fourier series, then instead of the single energy term $|c_1|^2$ we get a series of squares $|c_n|^2$ of Fourier coefficients c_n given by (6), Sec. 11.4. In this case we have a "**discrete spectrum**" (or "**point spectrum**") consisting of countably many isolated frequencies (infinitely many, in general), the corresponding $|c_n|^2$ being the contributions to the total energy.

Finally, a system whose solution can be represented by an integral (7) leads to the above integral for the energy, as is plausible from the cases just discussed.

Linearity. Fourier Transform of Derivatives

New transforms can be obtained from given ones by using

THEOREM 2

Linearity of the Fourier Transform

*The Fourier transform is a **linear operation**; that is, for any functions $f(x)$ and $g(x)$ whose Fourier transforms exist and any constants a and b, the Fourier transform of $af + bg$ exists, and*

(8)
$$\mathcal{F}(af + bg) = a\mathcal{F}(f) + b\mathcal{F}(g).$$

PROOF This is true because integration is a linear operation, so that (6) gives

$$\begin{aligned}\mathcal{F}\{af(x) + bg(x)\} &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} [af(x) + bg(x)]\, e^{-iwx}\, dx \\ &= a \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(x) e^{-iwx}\, dx + b \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x) e^{-iwx}\, dx \\ &= a\mathcal{F}\{f(x)\} + b\mathcal{F}\{g(x)\}.\end{aligned}$$

In applying the Fourier transform to differential equations, the key property is that differentiation of functions corresponds to multiplication of transforms by iw:

THEOREM 3

Fourier Transform of the Derivative of $f(x)$

Let $f(x)$ be continuous on the x-axis and $f(x) \to 0$ as $|x| \to \infty$. Furthermore, let $f'(x)$ be absolutely integrable on the x-axis. Then

(9)
$$\mathcal{F}\{f'(x)\} = iw\mathcal{F}\{f(x)\}.$$

PROOF From the definition of the Fourier transform we have

$$\mathcal{F}\{f'(x)\} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f'(x) e^{-iwx}\, dx.$$

Integrating by parts, we obtain

$$\mathcal{F}\{f'(x)\} = \frac{1}{\sqrt{2\pi}} \left[f(x) e^{-iwx} \Big|_{-\infty}^{\infty} - (-iw) \int_{-\infty}^{\infty} f(x) e^{-iwx}\, dx \right].$$

Since $f(x) \to 0$ as $|x| \to \infty$, the desired result follows, namely,

$$\mathcal{F}\{f'(x)\} = 0 + iw\mathcal{F}\{f(x)\}.$$

Two successive applications of (9) give

$$\mathcal{F}(f'') = iw\mathcal{F}(f') = (iw)^2\mathcal{F}(f).$$

Since $(iw)^2 = -w^2$, we have for the transform of the second derivative of f

$$\mathcal{F}\{f''(x)\} = -w^2\mathcal{F}\{f(x)\}. \tag{10}$$

Similarly for higher derivatives.

An application of (10) to differential equations will be given in Sec. 12.6. For the time being we show how (9) can be used to derive transforms.

EXAMPLE 3 **Application of the Operational Formula (9)**

Find the Fourier transform of xe^{-x^2} from Table III, Sec 11.10.

Solution. We use (9). By formula 9 in Table III

$$\begin{aligned}\mathcal{F}(xe^{-x^2}) &= \mathcal{F}\{-\tfrac{1}{2}(e^{-x^2})'\}\\ &= -\tfrac{1}{2}\mathcal{F}\{(e^{-x^2})'\}\\ &= -\tfrac{1}{2}iw\mathcal{F}(e^{-x^2})\\ &= -\frac{1}{2}iw\frac{1}{\sqrt{2}}e^{-w^2/4}\\ &= -\frac{iw}{2\sqrt{2}}e^{-w^2/4}.\end{aligned}$$

■

Convolution

The **convolution** $f * g$ of functions f and g is defined by

$$h(x) = (f * g)(x) = \int_{-\infty}^{\infty} f(p)g(x-p)\,dp = \int_{-\infty}^{\infty} f(x-p)g(p)\,dp. \tag{11}$$

The purpose is the same as in the case of Laplace transforms (Sec. 6.5): taking the convolution of two functions and then taking the transform of the convolution is the same as multiplying the transforms of these functions (and multiplying them by $\sqrt{2\pi}$):

THEOREM 4 **Convolution Theorem**

Suppose that $f(x)$ and $g(x)$ are piecewise continuous, bounded, and absolutely integrable on the x-axis. Then

$$\mathcal{F}(f * g) = \sqrt{2\pi}\,\mathcal{F}(f)\,\mathcal{F}(g). \tag{12}$$

PROOF By the definition,

$$\mathscr{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p) g(x-p)\, dp\, e^{-iwx}\, dx.$$

An interchange of the order of integration gives

$$\mathscr{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p) g(x-p) e^{-iwx}\, dx\, dp.$$

Instead of x we now take $x - p = q$ as a new variable of integration. Then $x = p + q$ and

$$\mathscr{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(p) g(q) e^{-iw(p+q)}\, dq\, dp.$$

This double integral can be written as a product of two integrals and gives the desired result

$$\mathscr{F}(f * g) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(p) e^{-iwp}\, dp \int_{-\infty}^{\infty} g(q) e^{-iwq}\, dq$$

$$= \frac{1}{\sqrt{2\pi}} [\sqrt{2\pi}\, \mathscr{F}(f)][\sqrt{2\pi}\, \mathscr{F}(g)] = \sqrt{2\pi}\, \mathscr{F}(f)\mathscr{F}(g).$$ ■

By taking the inverse Fourier transform on both sides of (12), writing $\hat{f} = \mathscr{F}(f)$ and $\hat{g} = \mathscr{F}(g)$ as before, and noting that $\sqrt{2\pi}$ and $1/\sqrt{2\pi}$ in (12) and (7) cancel each other, we obtain

(13) $$(f * g)(x) = \int_{-\infty}^{\infty} \hat{f}(w)\hat{g}(w) e^{iwx}\, dw,$$

a formula that will help us in solving partial differential equations (Sec. 12.6).

Discrete Fourier Transform (DFT), Fast Fourier Transform (FFT)

In using Fourier series, Fourier transforms, and trigonometric approximations (Sec. 11.6) we have to assume that a function $f(x)$, to be developed or transformed, is given on some interval, over which we integrate in the Euler formulas, etc. Now very often a function $f(x)$ is given only in terms of values at finitely many points, and one is interested in extending Fourier analysis to this case. The main application of such a "discrete Fourier analysis" concerns large amounts of equally spaced data, as they occur in telecommunication, time series analysis, and various simulation problems. In these situations, dealing with sampled values rather than with functions, we can replace the Fourier transform by the so-called **discrete Fourier transform (DFT)** as follows.

Let $f(x)$ be periodic, for simplicity of period 2π. We assume that N measurements of $f(x)$ are taken over the interval $0 \leqq x \leqq 2\pi$ at regularly spaced points

$$x_k = \frac{2\pi k}{N}, \qquad k = 0, 1, \cdots, N-1. \tag{14}$$

We also say that $f(x)$ is being **sampled** at these points. We now want to determine a **complex trigonometric polynomial**

$$q(x) = \sum_{n=0}^{N-1} c_n e^{inx_k} \tag{15}$$

that **interpolates** $f(x)$ at the nodes (14), that is, $q(x_k) = f(x_k)$, written out, with f_k denoting $f(x_k)$,

$$f_k = f(x_k) = q(x_k) = \sum_{n=0}^{N-1} c_n e^{inx_k}, \qquad k = 0, 1, \cdots, N-1. \tag{16}$$

Hence we must determine the coefficients $c_0, \cdots, c_{N-1}$ such that (16) holds. We do this by an idea similar to that in Sec. 11.1 for deriving the Fourier coefficients by using the orthogonality of the trigonometric system. Instead of integrals we now take sums. Namely, we multiply (16) by e^{-imx_k} (note the minus!) and sum over k from 0 to $N-1$. Then we interchange the order of the two summations and insert x_k from (14). This gives

$$\sum_{k=0}^{N-1} f_k e^{-imx_k} = \sum_{k=0}^{N-1}\sum_{n=0}^{N-1} c_n e^{i(n-m)x_k} = \sum_{n=0}^{N-1} c_n \sum_{k=0}^{N-1} e^{i(n-m)2\pi k/N}. \tag{17}$$

Now

$$e^{i(n-m)2\pi k/N} = [e^{i(n-m)2\pi/N}]^k.$$

We donote $[\cdots]$ by r. For $n = m$ we have $r = e^0 = 1$. The sum of *these* terms over k equals N, the number of these terms. For $n \neq m$ we have $r \neq 1$ and by the formula for a geometric sum [(6) in Sec. 15.1 with $q = r$ and $n = N-1$]

$$\sum_{k=0}^{N-1} r^k = \frac{1 - r^N}{1 - r} = 0$$

because $r^N = 1$; indeed, since k, m, and n are integers,

$$r^N = e^{i(n-m)2\pi k} = \cos 2\pi k(n-m) + i \sin 2\pi k(n-m) = 1 + 0 = 1.$$

This shows that the right side of (17) equals $c_m N$. Writing n for m and dividing by N, we thus obtain the desired coefficient formula

$$c_n = \frac{1}{N}\sum_{k=0}^{N-1} f_k e^{-inx_k} \qquad f_k = f(x_k), \quad n = 0, 1, \cdots, N-1. \tag{18*}$$

Since computation of the c_n (by the fast Fourier transform, below) involves successive halfing of the problem size N, it is practical to drop the factor $1/N$ from c_n and define the

discrete Fourier transform of the given signal $\mathbf{f} = [f_0 \quad \cdots \quad f_{N-1}]^\mathsf{T}$ to be the vector $\hat{\mathbf{f}} = [\hat{f}_0 \quad \cdots \quad \hat{f}_{N-1}]$ with components

(18) $$\hat{f}_n = Nc_n = \sum_{k=0}^{N-1} f_k e^{-inx_k}, \qquad f_k = f(x_k), \quad n = 0, \cdots, N-1.$$

This is the frequency spectrum of the signal.

In vector notation, $\hat{\mathbf{f}} = \mathbf{F}_N \mathbf{f}$, where the $N \times N$ **Fourier matrix** $\mathbf{F}_N = [e_{nk}]$ has the entries [given in (18)]

(19) $$e_{nk} = e^{-inx_k} = e^{-2\pi ink/N} = w^{nk}, \qquad w = w_N = e^{-2\pi i/N},$$

where $n, k = 0, \cdots, N-1$.

EXAMPLE 4 Discrete Fourier Transform (DFT). Sample of $N = 4$ Values

Let $N = 4$ measurements (sample values) be given. Then $w = e^{-2\pi i/N} = e^{-\pi i/2} = -i$ and thus $w^{nk} = (-i)^{nk}$. Let the sample values be, say $\mathbf{f} = [0 \quad 1 \quad 4 \quad 9]^\mathsf{T}$. Then by (18) and (19),

(20) $$\hat{\mathbf{f}} = \mathbf{F}_4\mathbf{f} = \begin{bmatrix} w^0 & w^0 & w^0 & w^0 \\ w^0 & w^1 & w^2 & w^3 \\ w^0 & w^2 & w^4 & w^6 \\ w^0 & w^3 & w^6 & w^9 \end{bmatrix} \mathbf{f} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 4 \\ 9 \end{bmatrix} = \begin{bmatrix} 14 \\ -4+8i \\ -6 \\ -4-8i \end{bmatrix}.$$

From the first matrix in (20) it is easy to infer what $\mathbf{F}_N$ looks like for arbitrary N, which in practice may be 1000 or more, for reasons given below. ■

From the DFT (the frequency spectrum) $\hat{\mathbf{f}} = \mathbf{F}_N\mathbf{f}$ we can recreate the given signal $\hat{\mathbf{f}} = \mathbf{F}_N^{-1}\mathbf{f}$, as we shall now prove. Here $\mathbf{F}_N$ and its complex conjugate $\overline{\mathbf{F}}_N = \frac{1}{N}[\overline{w}^{nk}]$ satisfy

(21a) $$\overline{\mathbf{F}}_N\mathbf{F}_N = \mathbf{F}_N\overline{\mathbf{F}}_N = N\mathbf{I}$$

where $\mathbf{I}$ is the $N \times N$ unit matrix; hence $\mathbf{F}_N$ has the inverse

(21b) $$\mathbf{F}_N^{-1} = \frac{1}{N}\overline{\mathbf{F}}_N.$$

PROOF We prove (21). By the multiplication rule (row times column) the product matrix $\mathbf{G}_N = \overline{\mathbf{F}}_N\mathbf{F}_N = [g_{jk}]$ in (21a) has the entries $g_{jk} =$ *Row* j *of* $\overline{\mathbf{F}}_N$ *times Column* k *of* $\mathbf{F}_N$. That is, writing $W = \overline{w}^j w^k$, we prove that

$$\begin{aligned} g_{jk} &= (\overline{w}^j w^k)^0 + (\overline{w}^j w^k)^1 + \cdots + (\overline{w}^j w^k)^{N-1} \\ &= W^0 + W^1 + \cdots + W^{N-1} = \begin{cases} 0 & \text{if} \quad j \neq k \\ N & \text{if} \quad j = k. \end{cases} \end{aligned}$$

Indeed, when $j = k$, then $\overline{w}^k w^k = (\overline{w}w)^k = (e^{2\pi i/N}e^{-2\pi i/N})^k = 1^k = 1$, so that the sum of *these* N terms equals N; these are the diagonal entries of $\mathbf{G}_N$. Also, when $j \neq k$, then $W \neq 1$ and we have a geometric sum (whose value is given by (6) in Sec. 15.1 with $q = W$ and $n = N - 1$)

$$W^0 + W^1 + \cdots + W^{N-1} = \frac{1 - W^N}{1 - W} = 0$$

because $W^N = (\overline{w}^j w^k)^N = (e^{2\pi i})^j (e^{-2\pi i})^k = 1^j \cdot 1^k = 1$. ■

We have seen that $\hat{\mathbf{f}}$ is the frequency spectrum of the signal $f(x)$. Thus the components $\hat{f}_n$ of $\hat{\mathbf{f}}$ give a resolution of the 2π-periodic function $f(x)$ into simple (complex) harmonics. Here one should use only n's that are much smaller than $N/2$, to avoid **aliasing**. By this we mean the effect caused by sampling at too few (equally spaced) points, so that, for instance, in a motion picture, rotating wheels appear as rotating too slowly or even in the wrong sense. Hence in applications, N is usually large. But this poses a problem. Eq. (18) requires $O(N)$ operations for any particular n, hence $O(N^2)$ operations for, say, all $n < N/2$. Thus, already for 1000 sample points the straightforward calculation would involve millions of operations. However, this difficulty can be overcome by the so-called **fast Fourier transform (FFT)**, for which codes are readily available (e.g., in Maple). The FFT is a computational method for the DFT that needs only $O(N)\log_2 N$ operations instead of $O(N^2)$. It makes the DFT a practical tool for large N. Here one chooses $N = 2^p$ (p integer) and uses the special form of the Fourier matrix to break down the given problem into smaller problems. For instance, when $N = 1000$, those operations are reduced by a factor $1000/\log_2 1000 \approx 100$.

The breakdown produces two problems of size $M = N/2$. This breakdown is possible because for $N = 2M$ we have in (19)

$$w_N^2 = w_{2M}^2 = (e^{-2\pi i/N})^2 = e^{-4\pi i/(2M)} = e^{-2\pi i/(M)} = w_M.$$

The given vector $\mathbf{f} = [f_0 \ \cdots \ f_{N-1}]^\mathsf{T}$ is split into two vectors with M components each, namely, $\mathbf{f}_{\text{ev}} = [f_0 \ f_2 \ \cdots \ f_{N-2}]^\mathsf{T}$ containing the even components of $\mathbf{f}$, and $\mathbf{f}_{\text{od}} = [f_1 \ f_3 \ \cdots \ f_{N-1}]^\mathsf{T}$ containing the odd components of $\mathbf{f}$. For $\mathbf{f}_{\text{ev}}$ and $\mathbf{f}_{\text{od}}$ we determine the DFTs

$$\hat{\mathbf{f}}_{\text{ev}} = [\hat{f}_{\text{ev},0} \quad \hat{f}_{\text{ev},2} \quad \cdots \quad \hat{f}_{\text{ev},N-2}]^\mathsf{T} = \mathbf{F}_M \mathbf{f}_{\text{ev}}$$

and

$$\hat{\mathbf{f}}_{\text{od}} = [\hat{f}_{\text{od},1} \quad \hat{f}_{\text{od},3} \quad \cdots \quad \hat{f}_{\text{od},N-1}]^\mathsf{T} = \mathbf{F}_M \mathbf{f}_{\text{od}}$$

involving the same $M \times M$ matrix $\mathbf{F}_M$. From these vectors we obtain the components of the DFT of the given vector f by the formulas

$$\text{(22)} \qquad \begin{aligned} &\text{(a)} \quad \hat{f}_n = \hat{f}_{\text{ev},n} + w_N^n \hat{f}_{\text{od},n} \qquad n = 0, \cdots, M-1 \\ &\text{(b)} \quad \hat{f}_{n+M} = \hat{f}_{\text{ev},n} - w_N^n \hat{f}_{\text{od},n} \qquad n = 0, \cdots, M-1. \end{aligned}$$

For $N = 2^p$ this breakdown can be repeated $p - 1$ times in order to finally arrive at $N/2$ problems of size 2 each, so that the number of multiplications is reduced as indicated above.

We show the reduction from $N = 4$ to $M = N/2 = 2$ and then prove (22).

EXAMPLE 5 Fast Fourier Transform (FFT). Sample of $N = 4$ Values

When $N = 4$, then $w = w_N = -i$ as in Example 4 and $M = N/2 = 2$, hence $w = w_M = e^{-2\pi i/2} = e^{-\pi i} = -1$. Consequently,

$$\hat{\mathbf{f}}_{\text{ev}} = \begin{bmatrix} \hat{f}_0 \\ \hat{f}_2 \end{bmatrix} = \mathbf{F}_2\mathbf{f}_{\text{ev}} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f_0 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_0 + f_2 \\ f_0 - f_2 \end{bmatrix}$$

$$\hat{\mathbf{f}}_{\text{od}} = \begin{bmatrix} \hat{f}_1 \\ \hat{f}_3 \end{bmatrix} = \mathbf{F}_2\mathbf{f}_{\text{od}} = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_3 \end{bmatrix} = \begin{bmatrix} f_1 + f_3 \\ f_1 - f_3 \end{bmatrix}.$$

From this and (22a) we obtain

$$\hat{f}_0 = \hat{f}_{\text{ev},0} + w_N^0 \hat{f}_{\text{od},0} = (f_0 + f_2) + (f_1 + f_3) = f_0 + f_1 + f_2 + f_3$$
$$\hat{f}_1 = \hat{f}_{\text{ev},1} + w_N^1 \hat{f}_{\text{od},1} = (f_0 - f_2) - i(f_1 + f_3) = f_0 - if_1 - f_2 + if_3.$$

Similarly, by (22b),

$$\hat{f}_2 = \hat{f}_{\text{ev},0} - w_N^0 \hat{f}_{\text{od},0} = (f_0 + f_2) - (f_1 + f_3) = f_0 - f_1 + f_2 - f_3$$
$$\hat{f}_3 = \hat{f}_{\text{ev},1} - w_N^1 \hat{f}_{\text{od},1} = (f_0 - f_2) - (-i)(f_1 - f_3) = f_0 + if_1 - f_2 - if_3.$$

This agrees with Example 4, as can be seen by replacing 0, 1, 4, 9 with f_0, f_1, f_2, f_3. ■

We prove (22). From (18) and (19) we have for the components of the DFT

$$\hat{f}_n = \sum_{k=0}^{N-1} w_N^{kn} f_k.$$

Splitting into two sums of $M = N/2$ terms each gives

$$\hat{f}_n = \sum_{k=0}^{M-1} w_N^{2kn} f_{2k} + \sum_{k=0}^{M-1} w_N^{(2k+1)n} f_{2k+1}.$$

We now use $w_N^2 = w_M$ and pull out w_N^n from under the second sum, obtaining

$$\hat{f}_n = \sum_{k=0}^{M-1} w_M^{kn} f_{\text{ev},k} + w_N^n \sum_{k=0}^{M-1} w_M^{kn} f_{\text{od},k}. \tag{23}$$

The two sums are $f_{\text{ev},n}$ and $f_{\text{od},n}$, the components of the "half-size" transforms $\mathbf{Ff}_{\text{ev}}$ and $\mathbf{Ff}_{\text{od}}$.

Formula (22a) is the same as (23). In (22b) we have $n + M$ instead of n. This causes a sign changes in (23), namely $-w_N^n$ before the second sum because

$$w_N^M = e^{-2\pi i M/N} = e^{-2\pi i/2} = e^{-\pi i} = -1.$$

This gives the minus in (22b) and completes the proof. ■

PROBLEM SET 11.9

1. **Review in complex.** Show that $1/i = -i$, $e^{-ix} = \cos x - i \sin x$, $e^{ix} + e^{-ix} = 2\cos x$, $e^{ix} - e^{-ix} = 2i \sin x$, $e^{ikx} = \cos kx + i \sin kx$.

2–11 FOURIER TRANSFORMS BY INTEGRATION

Find the Fourier transform of $f(x)$ (without using Table III in Sec. 11.10). Show details.

2. $f(x) = \begin{cases} e^{2ix} & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

3. $f(x) = \begin{cases} 1 & \text{if } a < x < b \\ 0 & \text{otherwise} \end{cases}$

4. $f(x) = \begin{cases} e^{kx} & \text{if } x < 0 \quad (k > 0) \\ 0 & \text{if } x > 0 \end{cases}$

5. $f(x) = \begin{cases} e^{x} & \text{if } -a < x < a \\ 0 & \text{otherwise} \end{cases}$

6. $f(x) = e^{-|x|} \quad (-\infty < x < \infty)$

7. $f(x) = \begin{cases} x & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$

8. $f(x) = \begin{cases} xe^{-x} & \text{if } -1 < x < 0 \\ 0 & \text{otherwise} \end{cases}$

9. $f(x) = \begin{cases} |x| & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

10. $f(x) = \begin{cases} x & \text{if } -1 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

11. $f(x) = \begin{cases} -1 & \text{if } -1 < x < 0 \\ 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$

12–17 USE OF TABLE III IN SEC. 11.10. OTHER METHODS

12. Find $\mathscr{F}(f(x))$ for $f(x) = xe^{-x}$ if $x > 0$, $f(x) = 0$ if $x < 0$, by (9) in the text and formula 5 in Table III (with $a = 1$). *Hint.* Consider xe^{-x} and e^{-x}.
13. Obtain $\mathscr{F}(e^{-x^2/2})$ from Table III.
14. In Table III obtain formula 7 from formula 8.
15. In Table III obtain formula 1 from formula 2.
16. **TEAM PROJECT. Shifting (a)** Show that if $f(x)$ has a Fourier transform, so does $f(x - a)$, and $\mathscr{F}\{f(x - a)\} = e^{-iwa}\mathscr{F}\{f(x)\}$.

    (b) Using (a), obtain formula 1 in Table III, Sec. 11.10, from formula 2.

    (c) Shifting on the *w*-Axis. Show that if $\hat{f}(w)$ is the Fourier transform of $f(x)$, then $\hat{f}(w - a)$ is the Fourier transform of $e^{iax}f(x)$.

    (d) Using (c), obtain formula 7 in Table III from 1 and formula 8 from 2.
17. What could give you the idea to solve Prob. 11 by using the solution of Prob. 9 and formula (9) in the text? Would this work?

18–25 DISCRETE FOURIER TRANSFORM

18. Verify the calculations in Example 4 of the text.
19. Find the transform of a general signal $f = [f_1 \quad f_2 \quad f_3 \quad f_4]^\mathsf{T}$ of four values.
20. Find the inverse matrix in Example 4 of the text and use it to recover the given signal.
21. Find the transform (the frequency spectrum) of a general signal of two values $[f_1 \quad f_2]^\mathsf{T}$.
22. Recreate the given signal in Prob. 21 from the frequency spectrum obtained.
23. Show that for a signal of eight sample values, $w = e^{-i/4} = (1 - i)/\sqrt{2}$. Check by squaring.
24. Write the Fourier matrix $\mathbf{F}$ for a sample of eight values explicitly.
25. **CAS Problem.** Calculate the inverse of the 8×8 Fourier matrix. Transform a general sample of eight values and transform it back to the given data.

11.10 Tables of Transforms

Table I. Fourier Cosine Transforms

See (2) in Sec. 11.8.

	$f(x)$	$\hat{f}_c(w) = \mathcal{F}_c(f)$	
1	$\begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}} \frac{\sin aw}{w}$	
2	$x^{a-1} \quad (0 < a < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(a)}{w^a} \cos \frac{a\pi}{2}$	($\Gamma(a)$ see App. A3.1.)
3	$e^{-ax} \quad (a > 0)$	$\sqrt{\frac{2}{\pi}} \left(\frac{a}{a^2 + w^2}\right)$	
4	$e^{-x^2/2}$	$e^{-w^2/2}$	
5	$e^{-ax^2} \quad (a > 0)$	$\frac{1}{\sqrt{2a}} e^{-w^2/(4a)}$	
6	$x^n e^{-ax} \quad (a > 0)$	$\sqrt{\frac{2}{\pi}} \frac{n!}{(a^2 + w^2)^{n+1}} \operatorname{Re}(a + iw)^{n+1}$	Re = Real part
7	$\begin{cases} \cos x & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}} \left[\frac{\sin a(1-w)}{1-w} + \frac{\sin a(1+w)}{1+w}\right]$	
8	$\cos(ax^2) \quad (a > 0)$	$\frac{1}{\sqrt{2a}} \cos\left(\frac{w^2}{4a} - \frac{\pi}{4}\right)$	
9	$\sin(ax^2) \quad (a > 0)$	$\frac{1}{\sqrt{2a}} \cos\left(\frac{w^2}{4a} + \frac{\pi}{4}\right)$	
10	$\frac{\sin ax}{x} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}} (1 - u(w - a))$	(See Sec. 6.3.)
11	$\frac{e^{-x} \sin x}{x}$	$\frac{1}{\sqrt{2\pi}} \arctan \frac{2}{w^2}$	
12	$J_0(ax) \quad (a > 0)$	$\sqrt{\frac{2}{\pi}} \frac{1}{\sqrt{a^2 - w^2}} (1 - u(w - a))$	(See Secs. 5.5, 6.3.)

Table II. Fourier Sine Transforms

See (5) in Sec. 11.8.

	$f(x)$	$\hat{f}_s(w) = \mathcal{F}_s(f)$	
1	$\begin{cases} 1 & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}} \left[\frac{1 - \cos aw}{w} \right]$	
2	$1/\sqrt{x}$	$1/\sqrt{w}$	
3	$1/x^{3/2}$	$2\sqrt{w}$	
4	x^{a-1} $(0 < a < 1)$	$\sqrt{\frac{2}{\pi}} \frac{\Gamma(a)}{w^a} \sin \frac{a\pi}{2}$	($\Gamma(a)$ see App. A3.1.)
5	e^{-ax} $(a > 0)$	$\sqrt{\frac{2}{\pi}} \left(\frac{w}{a^2 + w^2} \right)$	
6	$\frac{e^{-ax}}{x}$ $(a > 0)$	$\sqrt{\frac{2}{\pi}} \arctan \frac{w}{a}$	
7	$x^n e^{-ax}$ $(a > 0)$	$\sqrt{\frac{2}{\pi}} \frac{n!}{(a^2 + w^2)^{n+1}} \operatorname{Im}(a + iw)^{n+1}$	Im = Imaginary part
8	$xe^{-x^2/2}$	$we^{-w^2/2}$	
9	xe^{-ax^2} $(a > 0)$	$\frac{w}{(2a)^{3/2}} e^{-w^2/4a}$	
10	$\begin{cases} \sin x & \text{if } 0 < x < a \\ 0 & \text{otherwise} \end{cases}$	$\frac{1}{\sqrt{2\pi}} \left[\frac{\sin a(1 - w)}{1 - w} - \frac{\sin a(1 + w)}{1 + w} \right]$	
11	$\frac{\cos ax}{x}$ $(a > 0)$	$\sqrt{\frac{\pi}{2}} u(w - a)$	(See Sec. 6.3.)
12	$\arctan \frac{2a}{x}$ $(a > 0)$	$\sqrt{2\pi} \frac{\sin aw}{w} e^{-aw}$	

Table III. Fourier Transforms

See (6) in Sec. 11.9.

	$f(x)$	$\hat{f}(w) = \mathscr{F}(f)$
1	$\begin{cases} 1 & \text{if } -b < x < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}} \frac{\sin bw}{w}$
2	$\begin{cases} 1 & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{-ibw} - e^{-icw}}{iw\sqrt{2\pi}}$
3	$\frac{1}{x^2 + a^2} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}} \frac{e^{-a\|w\|}}{a}$
4	$\begin{cases} x & \text{if } 0 < x < b \\ 2x - b & \text{if } b < x < 2b \\ 0 & \text{otherwise} \end{cases}$	$\frac{-1 + 2e^{ibw} - e^{-2ibw}}{\sqrt{2\pi}w^2}$
5	$\begin{cases} e^{-ax} & \text{if } x > 0 \\ 0 & \text{otherwise} \end{cases} \quad (a > 0)$	$\frac{1}{\sqrt{2\pi}(a + iw)}$
6	$\begin{cases} e^{ax} & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{e^{(a-iw)c} - e^{(a-iw)b}}{\sqrt{2\pi}(a - iw)}$
7	$\begin{cases} e^{iax} & \text{if } -b < x < b \\ 0 & \text{otherwise} \end{cases}$	$\sqrt{\frac{2}{\pi}} \frac{\sin b(w - a)}{w - a}$
8	$\begin{cases} e^{iax} & \text{if } b < x < c \\ 0 & \text{otherwise} \end{cases}$	$\frac{i}{\sqrt{2\pi}} \frac{e^{ib(a-w)} - e^{ic(a-w)}}{a - w}$
9	$e^{-ax^2} \quad (a > 0)$	$\frac{1}{\sqrt{2a}} e^{-w^2/4a}$
10	$\frac{\sin ax}{x} \quad (a > 0)$	$\sqrt{\frac{\pi}{2}} \text{ if } \|w\| < a; \quad 0 \text{ if } \|w\| > a$

CHAPTER 11 REVIEW QUESTIONS AND PROBLEMS

1. What is a Fourier series? A Fourier cosine series? A half-range expansion? Answer from memory.

2. What are the Euler formulas? By what very important idea did we obtain them?

3. How did we proceed from 2π-periodic to general-periodic functions?

4. Can a discontinuous function have a Fourier series? A Taylor series? Why are such functions of interest to the engineer?

5. What do you know about convergence of a Fourier series? About the Gibbs phenomenon?

6. The output of an ODE can oscillate several times as fast as the input. How come?

7. What is approximation by trigonometric polynomials? What is the minimum square error?

8. What is a Fourier integral? A Fourier sine integral? Give simple examples.

9. What is the Fourier transform? The discrete Fourier transform?

10. What are Sturm–Liouville problems? By what idea are they related to Fourier series?

11–20 FOURIER SERIES. In Probs. 11, 13, 16, 20 find the Fourier series of $f(x)$ as given over one period and sketch $f(x)$ and partial sums. In Probs. 12, 14, 15, 17–19 give answers, with reasons. Show your work detail.

11. $f(x) = \begin{cases} 0 & \text{if } -2 < x < 0 \\ 2 & \text{if } \quad 0 < x < 2 \end{cases}$

12. Why does the series in Prob. 11 have no cosine terms?

13. $f(x) = \begin{cases} 0 & \text{if } -1 < x < 0 \\ x & \text{if } \quad 0 < x < 1 \end{cases}$

14. What function does the series of the cosine terms in Prob. 13 represent? The series of the sine terms?

15. What function do the series of the cosine terms and the series of the sine terms in the Fourier series of e^x $(-5 < x < 5)$ represent?

16. $f(x) = |x| \quad (-\pi < x < \pi)$

17. Find a Fourier series from which you can conclude that $1 - 1/3 + 1/5 - 1/7 + - \cdots = \pi/4$.

18. What function and series do you obtain in Prob. 16 by (termwise) differentiation?

19. Find the half-range expansions of $f(x) = x$ $(0 < x < 1)$.

20. $f(x) = 3x^2 \quad (-\pi < x < \pi)$

21–22 GENERAL SOLUTION

Solve, $y'' + \omega^2 y = r(t)$, where $|\omega| \neq 0, 1, 2, \cdots, r(t)$ is 2π-periodic and

21. $r(t) = 3t^2 \; (-\pi < t < \pi)$

22. $r(t) = |t| \; (-\pi < t < \pi)$

23–25 MINIMUM SQUARE ERROR

23. Compute the minimum square error for $f(x) = x/\pi$ $(-\pi < x < \pi)$ and trigonometric polynomials of degree $N = 1, \cdots, 5$.

24. How does the minimum square error change if you multiply $f(x)$ by a constant k?

25. Same task as in Prob. 23, for $f(x) = |x|/\pi$ $(-\pi < x < \pi)$. Why is E^* now much smaller (by a factor 100, approximately!)?

26–30 FOURIER INTEGRALS AND TRANSFORMS

Sketch the given function and represent it as indicated. If you have a CAS, graph approximate curves obtained by replacing ∞ with finite limits; also look for Gibbs phenomena.

26. $f(x) = x + 1$ if $0 < x < 1$ and 0 otherwise; by the Fourier sine transform

27. $f(x) = x$ if $0 < x < 1$ and 0 otherwise; by the Fourier integral

28. $f(x) = kx$ if $a < x < b$ and 0 otherwise; by the Fourier transform

29. $f(x) = x$ if $1 < x < a$ and 0 otherwise; by the Fourier cosine transform

30. $f(x) = e^{-2x}$ if $x > 0$ and 0 otherwise; by the Fourier transform

SUMMARY OF CHAPTER 11

Fourier Analysis. Partial Differential Equations (PDEs)

Fourier series concern **periodic functions** $f(x)$ of period $p = 2L$, that is, by definition $f(x + p) = f(x)$ for all x and some fixed $p > 0$; thus, $f(x + np) = f(x)$ for any integer n. These series are of the form

$$f(x) = a_0 + \sum_{n=1}^{\infty} \left(a_n \cos \frac{n\pi}{L} x + b_n \sin \frac{n\pi}{L} x \right) \qquad \text{(1)} \quad \text{(Sec. 11.2)}$$

with coefficients, called the **Fourier coefficients** of $f(x)$, given by the Euler formulas (Sec. 11.2)

$$\text{(2)} \quad a_0 = \frac{1}{2L} \int_{-L}^{L} f(x)\, dx, \qquad a_n = \frac{1}{L} \int_{-L}^{L} f(x) \cos \frac{n\pi x}{L}\, dx$$

$$b_n = \frac{1}{L} \int_{-L}^{L} f(x) \sin \frac{n\pi x}{L}\, dx$$

where $n = 1, 2, \cdots$. For period 2π we simply have (Sec. 11.1)

$$\text{(1*)} \quad f(x) = a_0 + \sum_{n=1}^{\infty} (a_n \cos nx + b_n \sin nx)$$

with the *Fourier coefficients* of $f(x)$ (Sec. 11.1)

$$a_0 = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(x)\, dx, \quad a_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \cos nx\, dx, \quad b_n = \frac{1}{\pi} \int_{-\pi}^{\pi} f(x) \sin nx\, dx.$$

Fourier series are fundamental in connection with periodic phenomena, particularly in models involving differential equations (Sec. 11.3, Chap, 12). If $f(x)$ is even $[f(-x) = f(x)]$ or odd $[f(-x) = -f(x)]$, they reduce to **Fourier cosine** or **Fourier sine series**, respectively (Sec. 11.2). If $f(x)$ is given for $0 \leqq x \leqq L$ only, it has two **half-range expansions** of period $2L$, namely, a cosine and a sine series (Sec. 11.2).

The set of cosine and sine functions in (1) is called the **trigonometric system**. Its most basic property is its **orthogonality** on an interval of length $2L$; that is, for all integers m and $n \neq m$ we have

$$\int_{-L}^{L} \cos \frac{m\pi x}{L} \cos \frac{n\pi x}{L}\, dx = 0, \qquad \int_{-L}^{L} \sin \frac{m\pi x}{L} \sin \frac{n\pi x}{L}\, dx = 0$$

and for all integers m and n,

$$\int_{-L}^{L} \cos \frac{m\pi x}{L} \sin \frac{n\pi x}{L}\, dx = 0.$$

This orthogonality was crucial in deriving the Euler formulas (2).

Partial sums of Fourier series minimize the **square error** (Sec. 11.4).

Replacing the trigonometric system in (1) by other orthogonal systems first leads to ***Sturm–Liouville problems*** (Sec. 11.5), which are boundary value problems for ODEs. These problems are ***eigenvalue problems*** and as such involve a parameter λ that is often related to frequencies and energies. The solutions to Sturm–Liouville problems are called ***eigenfunctions***. Similar considerations lead to other orthogonal series such as ***Fourier–Legendre series*** and ***Fourier–Bessel series*** classified as ***generalized Fourier series*** (Sec. 11.6).

Ideas and techniques of Fourier series extend to nonperiodic functions $f(x)$ defined on the entire real line; this leads to the **Fourier integral**

$$f(x) = \int_0^\infty [A(w)\cos wx + B(w)\sin wx]\, dw \qquad \text{(Sec. 11.7)} \tag{3}$$

where

$$A(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\cos wv\, dv, \qquad B(w) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\sin wv\, dv \tag{4}$$

or, in complex form (Sec. 11.9),

$$f(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \hat{f}(w)e^{iwx}\, dw \qquad (i = \sqrt{-1}) \tag{5}$$

where

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} f(x)e^{-iwx}\, dx. \tag{6}$$

Formula (6) transforms $f(x)$ into its **Fourier transform** $\hat{f}(w)$, and (5) is the inverse transform.

Related to this are the **Fourier cosine transform** (Sec. 11.8)

$$\hat{f}_c(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\cos wx\, dx \tag{7}$$

and the **Fourier sine transform** (Sec. 11.8)

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}}\int_0^\infty f(x)\sin wx\, dx. \tag{8}$$

The ***discrete Fourier transform (DFT)*** and a practical method of computing it, called the ***fast Fourier transform (FFT)***, are discussed in Sec. 11.9.

CHAPTER 12

Partial Differential Equations (PDEs)

A PDE is an equation that contains one or more partial derivatives of an unknown function that depends on at least two variables. Usually one of these deals with time t and the remaining with space (spatial variable(s)). The most important PDEs are the wave equations that can model the vibrating string (Secs. 12.2, 12.3, 12.4, 12.12) and the vibrating membrane (Secs. 12.8, 12.9, 12.10), the heat equation for temperature in a bar or wire (Secs. 12.5, 12.6), and the Laplace equation for electrostatic potentials (Secs. 12.6, 12.10, 12.11). PDEs are very important in dynamics, elasticity, heat transfer, electromagnetic theory, and quantum mechanics. They have a much wider range of applications than ODEs, which can model only the simplest physical systems. Thus PDEs are subjects of many ongoing research and development projects.

Realizing that modeling with PDEs is more involved than modeling with ODEs, we take a gradual, well-planned approach to modeling with PDEs. To do this we carefully derive the PDE that models the phenomena, such as the one-dimensional wave equation for a vibrating elastic string (say a violin string) in Sec. 12.2, and then solve the PDE in a separate section, that is, Sec. 12.3. In a similar vein, we derive the heat equation in Sec. 12.5 and then solve and generalize it in Sec. 12.6.

We derive these PDEs from physics and consider methods for solving initial and boundary value problems, that is, methods of obtaining solutions which satisfy the conditions required by the physical situations. In Secs. 12.7 and 12.12 we show how PDEs can also be solved by Fourier and Laplace transform methods.

COMMENT. ***Numerics for PDEs*** is explained in Secs. 21.4–21.7, which, for greater teaching flexibility, is designed to be independent of the other sections on numerics in Part E.

Prerequisites: Linear ODEs (Chap. 2), Fourier series (Chap. 11).
Sections that may be omitted in a shorter course: 12.7, 12.10–12.12.
References and Answers to Problems: App. 1 Part C, App. 2.

12.1 Basic Concepts of PDEs

A **partial differential equation (PDE)** is an equation involving one or more partial derivatives of an (unknown) function, call it u, that depends on two or more variables, often time t and one or several variables in space. The order of the highest derivative is called the **order** of the PDE. Just as was the case for ODEs, second-order PDEs will be the most important ones in applications.

Just as for ordinary differential equations (ODEs) we say that a PDE is **linear** if it is of the first degree in the unknown function u and its partial derivatives. Otherwise we call it **nonlinear**. Thus, all the equations in Example 1 are linear. We call a *linear* PDE **homogeneous** if each of its terms contains either u or one of its partial derivatives. Otherwise we call the equation **nonhomogeneous**. Thus, (4) in Example 1 (with f not identically zero) is nonhomogeneous, whereas the other equations are homogeneous.

EXAMPLE 1 Important Second-Order PDEs

(1) $$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}$$ ***One-dimensional wave equation***

(2) $$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}$$ ***One-dimensional heat equation***

(3) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0$$ ***Two-dimensional Laplace equation***

(4) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = f(x, y)$$ ***Two-dimensional Poisson equation***

(5) $$\frac{\partial^2 u}{\partial t^2} = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$ ***Two-dimensional wave equation***

(6) $$\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} = 0$$ ***Three-dimensional Laplace equation***

Here c is a positive constant, t is time, x, y, z are Cartesian coordinates, and *dimension* is the number of these coordinates in the equation. ■

A **solution** of a PDE in some region R of the space of the independent variables is a function that has all the partial derivatives appearing in the PDE in some domain D (definition in Sec. 9.6) containing R, and satisfies the PDE everywhere in R.

Often one merely requires that the function is continuous on the boundary of R, has those derivatives in the interior of R, and satisfies the PDE in the interior of R. Letting R lie in D simplifies the situation regarding derivatives on the boundary of R, which is then the same on the boundary as it is in the interior of R.

In general, the totality of solutions of a PDE is very large. For example, the functions

(7) $$u = x^2 - y^2, \qquad u = e^x \cos y, \qquad u = \sin x \cosh y, \qquad u = \ln (x^2 + y^2)$$

which are entirely different from each other, are solutions of (3), as you may verify. We shall see later that the unique solution of a PDE corresponding to a given physical problem will be obtained by the use of ***additional conditions*** arising from the problem. For instance, this may be the condition that the solution u assume given values on the boundary of the region R ("**boundary conditions**"). Or, when time t is one of the variables, u (or $u_t = \partial u/\partial t$ or both) may be prescribed at $t = 0$ ("**initial conditions**").

We know that if an ODE is linear and homogeneous, then from known solutions we can obtain further solutions by superposition. For PDEs the situation is quite similar:

THEOREM 1 Fundamental Theorem on Superposition

*If u_1 and u_2 are solutions of a **homogeneous linear** PDE in some region R, then*

$$u = c_1 u_1 + c_2 u_2$$

with any constants c_1 and c_2 is also a solution of that PDE in the region R.

The simple proof of this important theorem is quite similar to that of Theorem 1 in Sec. 2.1 and is left to the student.

Verification of solutions in Probs. 2–13 proceeds as for ODEs. Problems 16–23 concern PDEs solvable like ODEs. To help the student with them, we consider two typical examples.

EXAMPLE 2 **Solving $u_{xx} - u = 0$ Like an ODE**

Find solutions u of the PDE $u_{xx} - u = 0$ depending on x and y.

Solution. Since no y-derivatives occur, we can solve this PDE like $u'' - u = 0$. In Sec. 2.2 we would have obtained $u = Ae^x + Be^{-x}$ with constant A and B. Here A and B may be functions of y, so that the answer is

$$u(x, y) = A(y)e^x + B(y)e^{-x}$$

with arbitrary functions A and B. We thus have a great variety of solutions. Check the result by differentiation. ∎

EXAMPLE 3 **Solving $u_{xy} = -u_x$ Like an ODE**

Find solutions $u = u(x, y)$ of this PDE.

Solution. Setting $u_x = p$, we have $p_y = -p$, $p_y/p = -1$, $\ln|p| = -y + \tilde{c}(x)$, $p = c(x)e^{-y}$ and by integration with respect to x,

$$u(x, y) = f(x)e^{-y} + g(y) \qquad \text{where} \qquad f(x) = \int c(x)\, dx,$$

here, $f(x)$ and $g(y)$ are arbitrary. ∎

PROBLEM SET 12.1

1. **Fundamental theorem.** Prove it for second-order PDEs in two and three independent variables. *Hint.* Prove it by substitution.

2–13 VERIFICATION OF SOLUTIONS

Verifiy (by substitution) that the given function is a solution of the PDE. Sketch or graph the solution as a surface in space.

2–5 Wave Equation (1) with suitable c

2. $u = x^2 + t^2$
3. $u = \cos 4t \sin 2x$
4. $u = \sin kct \cos kx$
5. $u = \sin at \sin bx$

6–9 Heat Equation (2) with suitable c

6. $u = e^{-t} \sin x$
7. $u = e^{-\omega^2 c^2 t} \cos \omega x$
8. $u = e^{-9t} \sin \omega x$
9. $u = e^{-\pi^2 t} \cos 25x$

10–13 Laplace Equation (3)

10. $u = e^x \cos y,\ e^x \sin y$
11. $u = \arctan(y/x)$
12. $u = \cos y \sinh x,\ \sin y \cosh x$
13. $u = x/(x^2 + y^2),\ y/(x^2 + y^2)$
14. **TEAM PROJECT. Verification of Solutions**

    (a) Wave equation. Verify that $u(x, t) = v(x + ct) + w(x - ct)$ with any twice differentiable functions v and w satisfies (1).

    (b) Poisson equation. Verify that each u satisfies (4) with $f(x, y)$ as indicated.

    $u = y/x$ $\qquad f = 2y/x^3$

    $u = \sin xy$ $\qquad f = (x^2 + y^2) \sin xy$

    $u = e^{x^2 - y^2}$ $\qquad f = 4(x^2 + y^2)e^{x^2 - y^2}$

    $u = 1/\sqrt{x^2 + y^2}$ $\qquad f = (x^2 + y^2)^{-3/2}$

    (c) Laplace equation. Verify that $u = 1/\sqrt{x^2 + y^2 + z^2}$ satisfies (6) and $u = \ln(x^2 + y^2)$ satisfies (3). Is $u = 1/\sqrt{x^2 + y^2}$ a solution of (3)? Of what Poisson equation?

    (d) Verify that u with any (sufficiently often differentiable) v and w satisfies the given PDE.

    $u = v(x) + w(y)$ $\qquad u_{xy} = 0$

    $u = v(x)w(y)$ $\qquad uu_{xy} = u_x u_y$

    $u = v(x + 2t) + w(x - 2t)$ $\qquad u_{tt} = 4u_{xx}$

15. **Boundary value problem.** Verify that the function $u(x, y) = a \ln(x^2 + y^2) + b$ satisfies Laplace's equation

(3) and determine a and b so that u satisfies the boundary conditions $u = 110$ on the circle $x^2 + y^2 = 1$ and $u = 0$ on the circle $x^2 + y^2 = 100$.

16–23 PDEs SOLVABLE AS ODEs

This happens if a PDE involves derivatives with respect to one variable only (or can be transformed to such a form), so that the other variable(s) can be treated as parameter(s). Solve for $u = u(x, y)$:

16. $u_{yy} = 0$

17. $u_{xx} + 16\pi^2 u = 0$

18. $25u_{yy} - 4u = 0$

19. $u_y + y^2 u = 0$

20. $2u_{xx} + 9u_x + 4u = -3 \cos x - 29 \sin x$

21. $u_{yy} + 6u_y + 13u = 4e^{3y}$

22. $u_{xy} = u_x$

23. $x^2 u_{xx} + 2xu_x - 2u = 0$

24. Surface of revolution. Show that the solutions $z = z(x, y)$ of $yz_x = xz_y$ represent surfaces of revolution. Give examples. *Hint.* Use polar coordinates r, θ and show that the equation becomes $z_\theta = 0$.

25. System of PDEs. Solve $u_{xx} = 0$, $u_{yy} = 0$

12.2 Modeling: Vibrating String, Wave Equation

In this section we model a vibrating string, which will lead to our first important PDE, that is, equation (3) which will then be solved in Sec. 12.3. *The student should pay very close attention to this delicate modeling process and detailed derivation starting from scratch,* as the skills learned can be applied to modeling other phenomena in general and in particular to modeling a vibrating membrane (Sec. 12.7).

We want to derive the PDE modeling small transverse vibrations of an elastic string, such as a violin string. We place the string along the x-axis, stretch it to length L, and fasten it at the ends $x = 0$ and $x = L$. We then distort the string, and at some instant, call it $t = 0$, we release it and allow it to vibrate. The problem is to determine the vibrations of the string, that is, to find its deflection $u(x, t)$ at any point x and at any time $t > 0$; see Fig. 286.

$u(x, t)$ will be the solution of a PDE that is the model of our physical system to be derived. This PDE should not be too complicated, so that we can solve it. Reasonable simplifying assumptions (just as for ODEs modeling vibrations in Chap. 2) are as follows.

Physical Assumptions

1. The mass of the string per unit length is constant ("homogeneous string"). The string is perfectly elastic and does not offer any resistance to bending.
2. The tension caused by stretching the string before fastening it at the ends is so large that the action of the gravitational force on the string (trying to pull the string down a little) can be neglected.
3. The string performs small transverse motions in a vertical plane; that is, every particle of the string moves strictly vertically and so that the deflection and the slope at every point of the string always remain small in absolute value.

Under these assumptions we may expect solutions $u(x, t)$ that describe the physical reality sufficiently well.

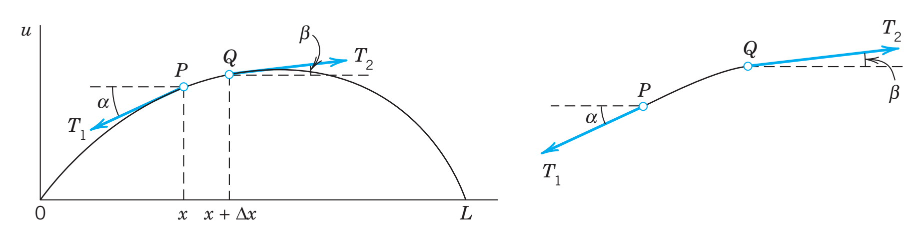


Fig. 286. Deflected string at fixed time t. Explanation on p. 544

Derivation of the PDE of the Model ("Wave Equation") from Forces

The model of the vibrating string will consist of a PDE ("wave equation") and additional conditions. To obtain the PDE, we consider the ***forces acting on a small portion of the string*** (Fig. 286). This method is typical of modeling in mechanics and elsewhere.

Since the string offers no resistance to bending, the tension is tangential to the curve of the string at each point. Let T_1 and T_2 be the tension at the endpoints P and Q of that portion. Since the points of the string move vertically, there is no motion in the horizontal direction. Hence the horizontal components of the tension must be constant. Using the notation shown in Fig. 286, we thus obtain

$$T_1 \cos \alpha = T_2 \cos \beta = T = \text{const.} \tag{1}$$

In the vertical direction we have two forces, namely, the vertical components $-T_1 \sin \alpha$ and $T_2 \sin \beta$ of T_1 and T_2; here the minus sign appears because the component at P is directed downward. By **Newton's second law** (Sec. 2.4) the resultant of these two forces is equal to the mass $\rho \Delta x$ of the portion times the acceleration $\partial^2 u/\partial t^2$, evaluated at some point between x and $x + \Delta x$; here ρ is the mass of the undeflected string per unit length, and Δx is the length of the portion of the undeflected string. (Δ is generally used to denote small quantities; this has nothing to do with the Laplacian ∇^2, which is sometimes also denoted by Δ.) Hence

$$T_2 \sin \beta - T_1 \sin \alpha = \rho \Delta x \frac{\partial^2 u}{\partial t^2}.$$

Using (1), we can divide this by $T_2 \cos \beta = T_1 \cos \alpha = T$, obtaining

$$\frac{T_2 \sin \beta}{T_2 \cos \beta} - \frac{T_1 \sin \alpha}{T_1 \cos \alpha} = \tan \beta - \tan \alpha = \frac{\rho \Delta x}{T} \frac{\partial^2 u}{\partial t^2}. \tag{2}$$

Now $\tan \alpha$ and $\tan \beta$ are the slopes of the string at x and $x + \Delta x$:

$$\tan \alpha = \left(\frac{\partial u}{\partial x}\right)\bigg|_x \qquad \text{and} \qquad \tan \beta = \left(\frac{\partial u}{\partial x}\right)\bigg|_{x+\Delta x}.$$

Here we have to write *partial* derivatives because u also depends on time t. Dividing (2) by Δx, we thus have

$$\frac{1}{\Delta x}\left[\left(\frac{\partial u}{\partial x}\right)\bigg|_{x+\Delta x} - \left(\frac{\partial u}{\partial x}\right)\bigg|_x\right] = \frac{\rho}{T} \frac{\partial^2 u}{\partial t^2}.$$

If we let Δx approach zero, we obtain the linear PDE

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad c^2 = \frac{T}{\rho}. \tag{3}$$

This is called the **one-dimensional wave equation**. We see that it is homogeneous and of the second order. The physical constant T/ρ is denoted by c^2 (instead of c) to indicate

that this constant is *positive,* a fact that will be essential to the form of the solutions. "One-dimensional" means that the equation involves only one space variable, x. In the next section we shall complete setting up the model and then show how to solve it by a general method that is probably the most important one for PDEs in engineering mathematics.

12.3 Solution by Separating Variables. Use of Fourier Series

We continue our work from Sec. 12.2, where we modeled a vibrating string and obtained the one-dimensional wave equation. We now have to complete the model by adding additional conditions and then solving the resulting model.

The model of a vibrating elastic string (a violin string, for instance) consists of the **one-dimensional wave equation**

$$\text{(1)} \qquad \frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2} \qquad c^2 = \frac{T}{\rho}$$

for the unknown deflection $u(x, t)$ of the string, a PDE that we have just obtained, and some ***additional conditions***, which we shall now derive.

Since the string is fastened at the ends $x = 0$ and $x = L$ (see Sec. 12.2), we have the two **boundary conditions**

$$\text{(2)} \qquad \text{(a)} \quad u(0, t) = 0, \qquad \text{(b)} \quad u(L, t) = 0, \qquad \text{for all } t \geqq 0.$$

Furthermore, the form of the motion of the string will depend on its *initial deflection* (deflection at time $t = 0$), call it $f(x)$, and on its *initial velocity* (velocity at $t = 0$), call it $g(x)$. We thus have the two **initial conditions**

$$\text{(3)} \qquad \text{(a)} \quad u(x, 0) = f(x), \qquad \text{(b)} \quad u_t(x, 0) = g(x) \qquad (0 \leqq x \leqq L)$$

where $u_t = \partial u/\partial t$. We now have to find a solution of the PDE (1) satisfying the conditions (2) and (3). This will be the solution of our problem. We shall do this in three steps, as follows.

Step 1. By the "**method of separating variables**" or *product method,* setting $u(x, t) = F(x)G(t)$, we obtain from (1) two ODEs, one for $F(x)$ and the other one for $G(t)$.

Step 2. We determine solutions of these ODEs that satisfy the boundary conditions (2).

Step 3. Finally, using **Fourier series**, we compose the solutions found in Step 2 to obtain a solution of (1) satisfying both (2) and (3), that is, the solution of our model of the vibrating string.

Step 1. Two ODEs from the Wave Equation (1)

In the **method of separating variables**, or *product method*, we determine solutions of the wave equation (1) of the form

$$\text{(4)} \qquad u(x, t) = F(x)G(t)$$

which are a product of two functions, each depending on only one of the variables x and t. This is a powerful general method that has various applications in engineering mathematics, as we shall see in this chapter. Differentiating (4), we obtain

$$\frac{\partial^2 u}{\partial t^2} = F\ddot{G} \qquad \text{and} \qquad \frac{\partial^2 u}{\partial x^2} = F''G$$

where dots denote derivatives with respect to t and primes derivatives with respect to x. By inserting this into the wave equation (1) we have

$$F\ddot{G} = c^2 F''G.$$

Dividing by c^2FG and simplifying gives

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F}.$$

The variables are now separated, the left side depending only on t and the right side only on x. Hence both sides must be constant because, if they were variable, then changing t or x would affect only one side, leaving the other unaltered. Thus, say,

$$\frac{\ddot{G}}{c^2 G} = \frac{F''}{F} = k.$$

Multiplying by the denominators gives immediately two ***ordinary*** DEs

(5) $$F'' - kF = 0$$

and

(6) $$\ddot{G} - c^2 kG = 0.$$

Here, the **separation constant** k is still arbitrary.

Step 2. Satisfying the Boundary Conditions (2)

We now determine solutions F and G of (5) and (6) so that $u = FG$ satisfies the boundary conditions (2), that is,

(7) $$u(0, t) = F(0)G(t) = 0, \qquad u(L, t) = F(L)G(t) = 0 \qquad \text{for all } t.$$

We first solve (5). If $G \equiv 0$, then $u = FG \equiv 0$, which is of no interest. Hence $G \not\equiv 0$ and then by (7),

(8) $$\text{(a)} \quad F(0) = 0, \qquad \text{(b)} \quad F(L) = 0.$$

We show that k must be negative. For $k = 0$ the general solution of (5) is $F = ax + b$, and from (8) we obtain $a = b = 0$, so that $F \equiv 0$ and $u = FG \equiv 0$, which is of no interest. For positive $k = \mu^2$ a general solution of (5) is

$$F = Ae^{\mu x} + Be^{-\mu x}$$

and from (8) we obtain $F \equiv 0$ as before (verify!). Hence we are left with the possibility of choosing k negative, say, $k = -p^2$. Then (5) becomes $F'' + p^2F = 0$ and has as a general solution

$$F(x) = A \cos px + B \sin px.$$

From this and (8) we have

$$F(0) = A = 0 \qquad \text{and then} \qquad F(L) = B \sin pL = 0.$$

We must take $B \neq 0$ since otherwise $F \equiv 0$. Hence $\sin pL = 0$. Thus

$$\text{(9)} \qquad pL = n\pi, \qquad \text{so that} \qquad p = \frac{n\pi}{L} \qquad (n \text{ integer}).$$

Setting $B = 1$, we thus obtain infinitely many solutions $F(x) = F_n(x)$, where

$$\text{(10)} \qquad F_n(x) = \sin \frac{n\pi}{L} x \qquad (n = 1, 2, \cdots).$$

These solutions satisfy (8). [For negative integer n we obtain essentially the same solutions, except for a minus sign, because $\sin(-\alpha) = -\sin \alpha$.]

We now solve (6) with $k = -p^2 = -(n\pi/L)^2$ resulting from (9), that is,

$$\text{(11*)} \qquad \ddot{G} + \lambda_n^2 G = 0 \qquad \text{where} \qquad \lambda_n = cp = \frac{cn\pi}{L}.$$

A general solution is

$$G_n(t) = B_n \cos \lambda_n t + B_n^* \sin \lambda_n t.$$

Hence solutions of (1) satisfying (2) are $u_n(x, t) = F_n(x)G_n(t) = G_n(t)F_n(x)$, written out

$$\text{(11)} \qquad u_n(x, t) = (B_n \cos \lambda_n t + B_n^* \sin \lambda_n t) \sin \frac{n\pi}{L} x \qquad (n = 1, 2, \cdots).$$

These functions are called the **eigenfunctions**, or *characteristic functions*, and the values $\lambda_n = cn\pi/L$ are called the **eigenvalues**, or *characteristic values*, of the vibrating string. The set $\{\lambda_1, \lambda_2, \cdots\}$ is called the **spectrum**.

Discussion of Eigenfunctions. We see that each u_n represents a harmonic motion having the **frequency** $\lambda_n/2\pi = cn/2L$ cycles per unit time. This motion is called the nth **normal mode** of the string. The first normal mode is known as the *fundamental mode* ($n = 1$), and the others are known as *overtones;* musically they give the octave, octave plus fifth, etc. Since in (11)

$$\sin \frac{n\pi x}{L} = 0 \qquad \text{at} \qquad x = \frac{L}{n}, \frac{2L}{n}, \cdots, \frac{n-1}{n}L,$$

the nth normal mode has $n - 1$ **nodes**, that is, points of the string that do not move (in addition to the fixed endpoints); see Fig. 287.

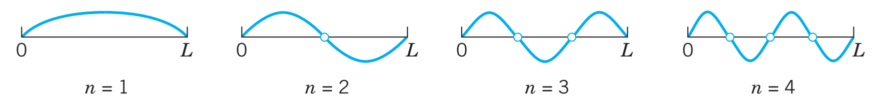


Fig. 287. Normal modes of the vibrating string

Figure 288 shows the second normal mode for various values of t. At any instant the string has the form of a sine wave. When the left part of the string is moving down, the other half is moving up, and conversely. For the other modes the situation is similar.

Tuning is done by changing the tension T. Our formula for the frequency $\lambda_n/2\pi = cn/2L$ of u_n with $c = \sqrt{T/\rho}$ [see (3), Sec. 12.2] confirms that effect because it shows that the frequency is proportional to the tension. T cannot be increased indefinitely, but can you see what to do to get a string with a high fundamental mode? (Think of both L and ρ.) Why is a violin smaller than a double-bass?

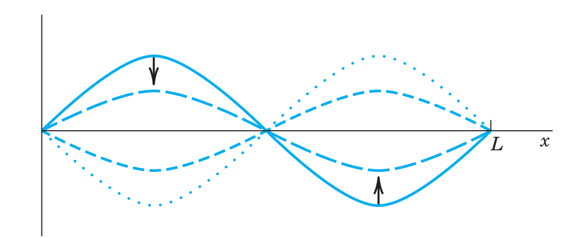


Fig. 288. Second normal mode for various values of t

Step 3. Solution of the Entire Problem. Fourier Series

The eigenfunctions (11) satisfy the wave equation (1) and the boundary conditions (2) (string fixed at the ends). A single u_n will generally not satisfy the initial conditions (3). But since the wave equation (1) is linear and homogeneous, it follows from Fundamental Theorem 1 in Sec. 12.1 that the sum of finitely many solutions u_n is a solution of (1). To obtain a solution that also satisfies the initial conditions (3), we consider the infinite series (with $\lambda_n = cn\pi/L$ as before)

$$u(x, t) = \sum_{n=1}^{\infty} u_n(x, t) = \sum_{n=1}^{\infty} (B_n \cos \lambda_n t + B_n^* \sin \lambda_n t) \sin \frac{n\pi}{L} x. \tag{12}$$

Satisfying Initial Condition (3a) (Given Initial Displacement). From (12) and (3a) we obtain

$$u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi}{L} x = f(x). \qquad (0 \leqq x \leqq L). \tag{13}$$

Hence we must choose the B_n's so that $u(x, 0)$ becomes the **Fourier sine series** of $f(x)$. Thus, by (4) in Sec. 11.3,

$$B_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} \, dx, \qquad n = 1, 2, \cdots. \tag{14}$$

Satisfying Initial Condition (3b) (Given Initial Velocity). Similarly, by differentiating (12) with respect to t and using (3b), we obtain

$$\frac{\partial u}{\partial t}\bigg|_{t=0} = \left[\sum_{n=1}^{\infty} (-B_n\lambda_n \sin \lambda_n t + B_n^*\lambda_n \cos \lambda_n t) \sin \frac{n\pi x}{L}\right]_{t=0}$$

$$= \sum_{n=1}^{\infty} B_n^*\lambda_n \sin \frac{n\pi x}{L} = g(x).$$

Hence we must choose the B_n^*'s so that for $t = 0$ the derivative $\partial u/\partial t$ becomes the Fourier sine series of $g(x)$. Thus, again by (4) in Sec. 11.3,

$$B_n^*\lambda_n = \frac{2}{L}\int_0^L g(x) \sin \frac{n\pi x}{L}\, dx.$$

Since $\lambda_n = cn\pi/L$, we obtain by division

(15) $$B_n^* = \frac{2}{cn\pi}\int_0^L g(x) \sin \frac{n\pi x}{L}\, dx, \qquad n = 1, 2, \cdots.$$

Result. Our discussion shows that $u(x, t)$ given by (12) with coefficients (14) and (15) is a solution of (1) that satisfies all the conditions in (2) and (3), provided the series (12) converges and so do the series obtained by differentiating (12) twice termwise with respect to x and t and have the sums $\partial^2 u/\partial x^2$ and $\partial^2 u/\partial t^2$, respectively, which are continuous.

Solution (12) Established. According to our derivation, the solution (12) is at first a purely formal expression, but we shall now establish it. For the sake of simplicity we consider only the case when the initial velocity $g(x)$ is identically zero. Then the B_n^* are zero, and (12) reduces to

(16) $$u(x, t) = \sum_{n=1}^{\infty} B_n \cos \lambda_n t \sin \frac{n\pi x}{L}, \qquad \lambda_n = \frac{cn\pi}{L}.$$

It is possible to ***sum this series***, that is, to write the result in a closed or finite form. For this purpose we use the formula [see (11), App. A3.1]

$$\cos \frac{cn\pi}{L} t \sin \frac{n\pi}{L} x = \frac{1}{2}\left[\sin\left\{\frac{n\pi}{L}(x - ct)\right\} + \sin\left\{\frac{n\pi}{L}(x + ct)\right\}\right].$$

Consequently, we may write (16) in the form

$$u(x, t) = \frac{1}{2}\sum_{n=1}^{\infty} B_n \sin\left\{\frac{n\pi}{L}(x - ct)\right\} + \frac{1}{2}\sum_{n=1}^{\infty} B_n \sin\left\{\frac{n\pi}{L}(x + ct)\right\}.$$

These two series are those obtained by substituting $x - ct$ and $x + ct$, respectively, for the variable x in the Fourier sine series (13) for $f(x)$. Thus

(17) $$u(x, t) = \tfrac{1}{2}[f^*(x - ct) + f^*(x + ct)]$$

where f^* is the odd periodic extension of f with the period $2L$ (Fig. 289). Since the initial deflection $f(x)$ is continuous on the interval $0 \leqq x \leqq L$ and zero at the endpoints, it follows from (17) that $u(x, t)$ is a continuous function of both variables x and t for all values of the variables. By differentiating (17) we see that $u(x, t)$ is a solution of (1), provided $f(x)$ is twice differentiable on the interval $0 < x < L$, and has one-sided second derivatives at $x = 0$ and $x = L$, which are zero. Under these conditions $u(x, t)$ is established as a solution of (1), satisfying (2) and (3) with $g(x) \equiv 0$. ■

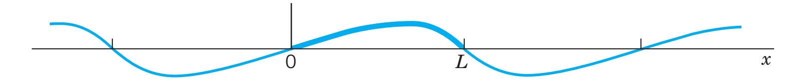


Fig. 289. Odd periodic extension of $f(x)$

Generalized Solution. If $f'(x)$ and $f''(x)$ are merely piecewise continuous (see Sec. 6.1), or if those one-sided derivatives are not zero, then for each t there will be finitely many values of x at which the second derivatives of u appearing in (1) do not exist. Except at these points the wave equation will still be satisfied. We may then regard $u(x, t)$ as a "**generalized solution**," as it is called, that is, as a solution in a broader sense. For instance, a triangular initial deflection as in Example 1 (below) leads to a generalized solution.

Physical Interpretation of the Solution (17). The graph of $f^*(x - ct)$ is obtained from the graph of $f^*(x)$ by shifting the latter ct units to the right (Fig. 290). This means that $f^*(x - ct)(c > 0)$ represents a wave that is traveling to the right as t increases. Similarly, $f^*(x + ct)$ represents a wave that is traveling to the left, and $u(x, t)$ is the superposition of these two waves.

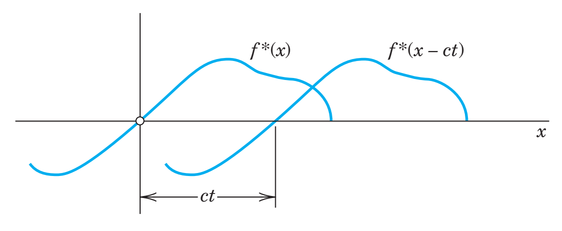


Fig. 290. Interpretation of (17)

EXAMPLE 1 Vibrating String if the Initial Deflection Is Triangular

Find the solution of the wave equation (1) satisfying (2) and corresponding to the triangular initial deflection

$$f(x) = \begin{cases} \dfrac{2k}{L}x & \text{if} \quad 0 < x < \dfrac{L}{2} \\[2ex] \dfrac{2k}{L}(L - x) & \text{if} \quad \dfrac{L}{2} < x < L \end{cases}$$

and initial velocity zero. (Figure 291 shows $f(x) = u(x, 0)$ at the top.)

Solution. Since $g(x) \equiv 0$, we have $B_n^* = 0$ in (12), and from Example 4 in Sec. 11.3 we see that the B_n are given by (5), Sec. 11.3. Thus (12) takes the form

$$u(x, t) = \frac{8k}{\pi^2}\left[\frac{1}{1^2}\sin\frac{\pi}{L}x\cos\frac{\pi c}{L}t - \frac{1}{3^2}\sin\frac{3\pi}{L}x\cos\frac{3\pi c}{L}t + - \cdots\right].$$

For graphing the solution we may use $u(x, 0) = f(x)$ and the above interpretation of the two functions in the representation (17). This leads to the graph shown in Fig. 291. ■

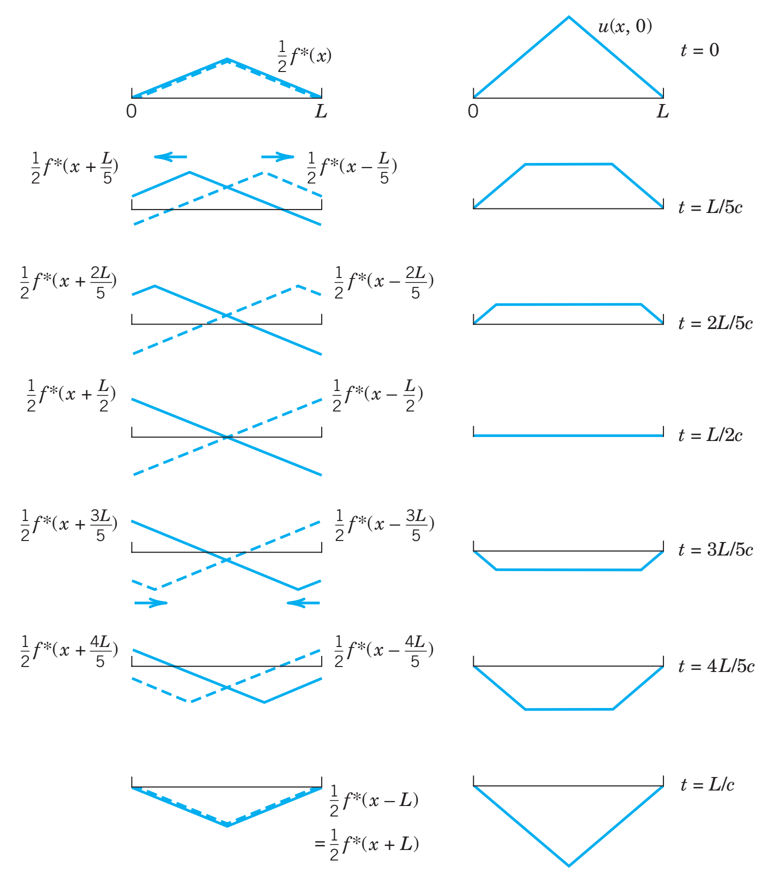


Fig. 291. Solution $u(x, t)$ in Example 1 for various values of t (right part of the figure) obtained as the superposition of a wave traveling to the right (dashed) and a wave traveling to the left (left part of the figure)

PROBLEM SET 12.3

1. **Frequency.** How does the frequency of the fundamental mode of the vibrating string depend on the length of the string? On the mass per unit length? What happens if we double the tension? Why is a contrabass larger than a violin?

2. **Physical Assumptions.** How would the motion of the string change if Assumption 3 were violated? Assumption 2? The second part of Assumption 1? The first part? Do we really need all these assumptions?

3. **String of length π.** Write down the derivation in this section for length $L = \pi$, to see the very substantial simplification of formulas in this case that may show ideas more clearly.

4. **CAS PROJECT. Graphing Normal Modes.** Write a program for graphing u_n with $L = \pi$ and c^2 of your choice similarly as in Fig. 287. Apply the program to u_2, u_3, u_4. Also graph these solutions as surfaces over the xt-plane. Explain the connection between these two kinds of graphs.

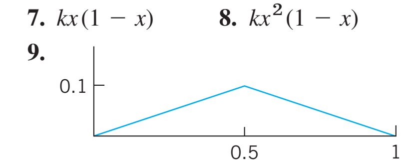
5–13 DEFLECTION OF THE STRING

Find $u(x, t)$ for the string of length $L = 1$ and $c^2 = 1$ when the initial velocity is zero and the initial deflection with small k (say, 0.01) is as follows. Sketch or graph $u(x, t)$ as in Fig. 291 in the text.

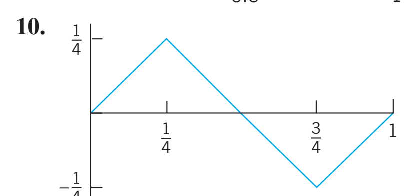
5. $k \sin 3\pi x$
6. $k(\sin \pi x - \frac{1}{2}\sin 2\pi x)$

7. $kx(1 - x)$ **8.** $kx^2(1 - x)$

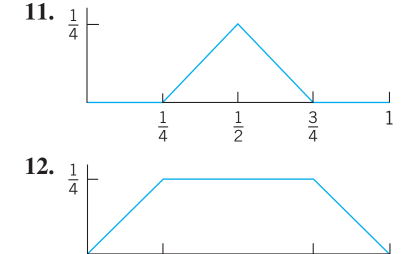
9.



10.



11.



12.

13. $2x - 4x^2$ if $0 < x < \frac{1}{2}$, 0 if $\frac{1}{2} < x < 1$

14. Nonzero initial velocity. Find the deflection $u(x, t)$ of the string of length $L = \pi$ and $c^2 = 1$ for zero initial displacement and "triangular" initial velocity $u_t(x, 0) = 0.01x$ if $0 \leqq x \leqq \frac{1}{2}\pi$, $u_t(x, 0) = 0.01(\pi - x)$ if $\frac{1}{2}\pi \leqq x \leqq \pi$. (Initial conditions with $u_t(x, 0) \neq 0$ are hard to realize experimentally.)

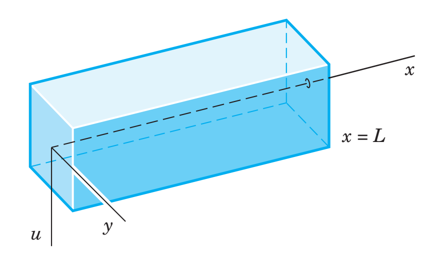


Fig. 292. Elastic beam

15–20 SEPARATION OF A FOURTH-ORDER PDE. VIBRATING BEAM

By the principles used in modeling the string it can be shown that small free vertical vibrations of a uniform elastic beam (Fig. 292) are modeled by the fourth-order PDE

$$\frac{\partial^2 u}{\partial t^2} = -c^2 \frac{\partial^4 u}{\partial x^4} \qquad (21) \qquad \text{(Ref. [C11])}$$

where $c^2 = EI/\rho A$ (E = Young's modulus of elasticity, I = moment of intertia of the cross section with respect to the y-axis in the figure, ρ = density, A = cross-sectional area). (*Bending* of a beam under a load is discussed in Sec. 3.3.)

15. Substituting $u = F(x)G(t)$ into (21), show that

$$F^{(4)}/F = -\ddot{G}/c^2 G = \beta^4 = \text{const},$$
$$F(x) = A \cos \beta x + B \sin \beta x + C \cosh \beta x + D \sinh \beta x,$$
$$G(t) = a \cos c\beta^2 t + b \sin c\beta^2 t.$$

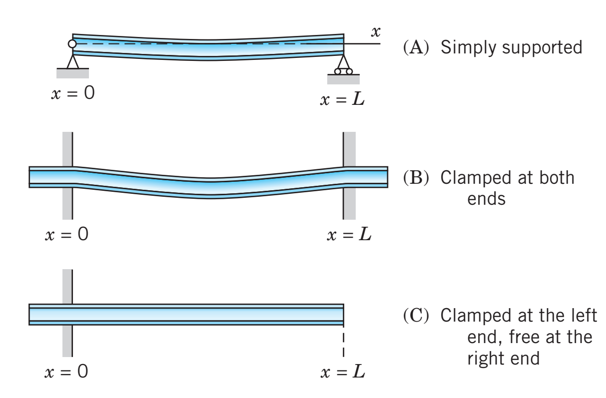


Fig. 293. Supports of a beam

16. Simply supported beam in Fig. 293A. Find solutions $u_n = F_n(x)G_n(t)$ of (21) corresponding to zero initial velocity and satisfying the boundary conditions (see Fig. 293A)

$$u(0, t) = 0, \; u(L, t) = 0$$
(ends simply supported for all times t),
$$u_{xx}(0, t) = 0, \; u_{xx}(L, t) = 0$$
(zero moments, hence zero curvature, at the ends).

17. Find the solution of (21) that satisfies the conditions in Prob. 16 as well as the initial condition

$$u(x, 0) = f(x) = x(L - x).$$

18. Compare the results of Probs. 17 and 7. What is the basic difference between the frequencies of the normal modes of the vibrating string and the vibrating beam?

19. Clamped beam in Fig. 293B. What are the boundary conditions for the clamped beam in Fig. 293B? Show that F in Prob. 15 satisfies these conditions if βL is a solution of the equation

$$\cosh \beta L \cos \beta L = 1. \qquad (22)$$

Determine approximate solutions of (22), for instance, graphically from the intersections of the curves of $\cos \beta L$ and $1/\cosh \beta L$.

20. Clamped-free beam in Fig. 293C. If the beam is clamped at the left and free at the right (Fig. 293C), the boundary conditions are

$$u(0, t) = 0, \qquad u_x(0, t) = 0,$$
$$u_{xx}(L, t) = 0, \qquad u_{xxx}(L, t) = 0.$$

Show that F in Prob. 15 satisfies these conditions if βL is a solution of the equation

$$\cosh \beta L \cos \beta L = -1. \tag{23}$$

Find approximate solutions of (23).

12.4 D'Alembert's Solution of the Wave Equation. Characteristics

It is interesting that the solution (17), Sec. 12.3, of the wave equation

$$\frac{\partial^2 u}{\partial t^2} = c^2 \frac{\partial^2 u}{\partial x^2}, \qquad c^2 = \frac{T}{\rho}, \tag{1}$$

can be immediately obtained by transforming (1) in a suitable way, namely, by introducing the new independent variables

$$v = x + ct, \qquad w = x - ct. \tag{2}$$

Then u becomes a function of v and w. The derivatives in (1) can now be expressed in terms of derivatives with respect to v and w by the use of the chain rule in Sec. 9.6. Denoting partial derivatives by subscripts, we see from (2) that $v_x = 1$ and $w_x = 1$. For simplicity let us denote $u(x, t)$, as a function of v and w, by the same letter u. Then

$$u_x = u_v v_x + u_w w_x = u_v + u_w.$$

We now apply the chain rule to the right side of this equation. We assume that all the partial derivatives involved are continuous, so that $u_{wv} = u_{vw}$. Since $v_x = 1$ and $w_x = 1$, we obtain

$$u_{xx} = (u_v + u_w)_x = (u_v + u_w)_v v_x + (u_v + u_w)_w w_x = u_{vv} + 2u_{vw} + u_{ww}.$$

Transforming the other derivative in (1) by the same procedure, we find

$$u_{tt} = c^2(u_{vv} - 2u_{vw} + u_{ww}).$$

By inserting these two results in (1) we get (see footnote 2 in App. A3.2)

$$u_{vw} \equiv \frac{\partial^2 u}{\partial w\, \partial v} = 0. \tag{3}$$

The point of the present method is that (3) can be readily solved by two successive integrations, first with respect to w and then with respect to v. This gives

$$\frac{\partial u}{\partial v} = h(v) \qquad \text{and} \qquad u = \int h(v)\, dv + \psi(w).$$

Here $h(v)$ and $\psi(w)$ are arbitrary functions of v and w, respectively. Since the integral is a function of v, say, $\phi(v)$, the solution is of the form $u = \phi(v) + \psi(w)$. In terms of x and t, by (2), we thus have

$$u(x, t) = \phi(x + ct) + \psi(x - ct). \tag{4}$$

This is known as **d'Alembert's solution**[1] of the wave equation (1).

Its derivation was much more elegant than the method in Sec. 12.3, but d'Alembert's method is special, whereas the use of Fourier series applies to various equations, as we shall see.

D'Alembert's Solution Satisfying the Initial Conditions

$$\text{(a)} \quad u(x, 0) = f(x), \qquad \text{(b)} \quad u_t(x, 0) = g(x). \tag{5}$$

These are the same as (3) in Sec. 12.3. By differentiating (4) we have

$$u_t(x, t) = c\phi'(x + ct) - c\psi'(x - ct) \tag{6}$$

where primes denote derivatives with respect to the *entire* arguments $x + ct$ and $x - ct$, respectively, and the minus sign comes from the chain rule. From (4)–(6) we have

$$u(x, 0) = \phi(x) + \psi(x) = f(x), \tag{7}$$

$$u_t(x, 0) = c\phi'(x) + c\psi'(x) = g(x). \tag{8}$$

Dividing (8) by c and integrating with respect to x, we obtain

$$\phi(x) - \psi(x) = k(x_0) + \frac{1}{c}\int_{x_0}^{x} g(s)\, ds, \qquad k(x_0) = \phi(x_0) - \psi(x_0). \tag{9}$$

If we add this to (7), then ψ drops out and division by 2 gives

$$\phi(x) = \frac{1}{2} f(x) + \frac{1}{2c}\int_{x_0}^{x} g(s)\, ds + \frac{1}{2} k(x_0). \tag{10}$$

Similarly, subtraction of (9) from (7) and division by 2 gives

$$\psi(x) = \frac{1}{2} f(x) - \frac{1}{2c}\int_{x_0}^{x} g(s)\, ds - \frac{1}{2} k(x_0). \tag{11}$$

In (10) we replace x by $x + ct$; we then get an integral from x_0 to $x + ct$. In (11) we replace x by $x - ct$ and get minus an integral from x_0 to $x - ct$ or plus an integral from $x - ct$ to x_0. Hence addition of $\phi(x + ct)$ and $\psi(x - ct)$ gives $u(x, t)$ [see (4)] in the form

$$u(x, t) = \frac{1}{2}[f(x + ct) + f(x - ct)] + \frac{1}{2c}\int_{x-ct}^{x+ct} g(s)\, ds. \tag{12}$$

---

[1]JEAN LE ROND D'ALEMBERT (1717–1783), French mathematician, also known for his important work in mechanics.

We mention that the general theory of PDEs provides a systematic way for finding the transformation (2) that simplifies (1). See Ref. [C8] in App. 1.

If the initial velocity is zero, we see that this reduces to

$$u(x, t) = \tfrac{1}{2}[f(x + ct) + f(x - ct)], \tag{13}$$

in agreement with (17) in Sec. 12.3. You may show that because of the boundary conditions (2) in that section the function f must be odd and must have the period $2L$.

Our result shows that the two initial conditions [the functions $f(x)$ and $g(x)$ in (5)] determine the solution uniquely.

The solution of the wave equation by the Laplace transform method will be shown in Sec. 12.11.

Characteristics. Types and Normal Forms of PDEs

The idea of d'Alembert's solution is just a special instance of the **method of characteristics**. This concerns PDEs of the form

$$Au_{xx} + 2Bu_{xy} + Cu_{yy} = F(x, y, u, u_x, u_y) \tag{14}$$

(as well as PDEs in more than two variables). Equation (14) is called **quasilinear** because it is linear in the highest derivatives (but may be arbitrary otherwise). There are three types of PDEs (14), depending on the discriminant $AC - B^2$, as follows.

Type	Defining Condition	Example in Sec. 12.1
Hyperbolic	$AC - B^2 < 0$	Wave equation (1)
Parabolic	$AC - B^2 = 0$	Heat equation (2)
Elliptic	$AC - B^2 > 0$	Laplace equation (3)

Note that (1) and (2) in Sec. 12.1 involve t, but to have y as in (14), we set $y = ct$ in (1), obtaining $u_{tt} - c^2u_{xx} = c^2(u_{yy} - u_{xx}) = 0$. And in (2) we set $y = c^2t$, so that $u_t - c^2u_{xx} = c^2(u_y - u_{xx})$.

A, B, C may be functions of x, y, so that a PDE may be **of mixed type**, that is, of different type in different regions of the xy-plane. An important mixed-type PDE is the **Tricomi equation** (see Prob. 10).

Transformation of (14) to Normal Form. The normal forms of (14) and the corresponding transformations depend on the type of the PDE. They are obtained by solving the **characteristic equation** of (14), which is the ODE

$$Ay'^2 - 2By' + C = 0 \tag{15}$$

where $y' = dy/dx$ (note $-2B$, not $+2B$). The solutions of (15) are called the **characteristics** of (14), and we write them in the form $\Phi(x, y) = \text{const}$ and $\Psi(x, y) = \text{const}$. Then the transformations giving new variables v, w instead of x, y and the normal forms of (14) are as follows.

Type	New Variables		Normal Form
Hyperbolic	$v = \Phi$	$w = \Psi$	$u_{vw} = F_1$
Parabolic	$v = x$	$w = \Phi = \Psi$	$u_{ww} = F_2$
Elliptic	$v = \frac{1}{2}(\Phi + \Psi)$	$w = \frac{1}{2i}(\Phi - \Psi)$	$u_{vv} + u_{ww} = F_3$

Here, $\Phi = \Phi(x, y), \Psi = \Psi(x, y), F_1 = F_1(v, w, u, u_v, u_w)$, etc., and we denote u as function of v, w again by u, for simplicity. We see that the normal form of a hyperbolic PDE is as in d'Alembert's solution. In the parabolic case we get just one family of solutions $\Phi = \Psi$. In the elliptic case, $i = \sqrt{-1}$, and the characteristics are complex and are of minor interest. For derivation, see Ref. [GenRef3] in App. 1.

EXAMPLE 1 D'Alembert's Solution Obtained Systematically

The theory of characteristics gives d'Alembert's solution in a systematic fashion. To see this, we write the wave equation $u_{tt} - c^2 u_{xx} = 0$ in the form (14) by setting $y = ct$. By the chain rule, $u_t = u_y y_t = cu_y$ and $u_{tt} = c^2 u_{yy}$. Division by c^2 gives $u_{xx} - u_{yy} = 0$, as stated before. Hence the characteristic equation is $y'^2 - 1 = (y' + 1)(y' - 1) = 0$. The two families of solutions (characteristics) are $\Phi(x, y) = y + x = \text{const}$ and $\Psi(x, y) = y - x = \text{const}$. This gives the new variables $v = \Phi = y + x = ct + x$ and $w = \Psi = y - x = ct - x$ and d'Alembert's solution $u = f_1(x + ct) + f_2(x - ct)$. ■

PROBLEM SET 12.4

1. Show that c is the speed of each of the two waves given by (4).
2. Show that, because of the boundary conditions (2), Sec. 12.3, the function f in (13) of this section must be odd and of period $2L$.
3. If a steel wire 2 m in length weighs 0.9 nt (about 0.20 lb) and is stretched by a tensile force of 300 nt (about 67.4 lb), what is the corresponding speed of transverse waves?
4. What are the frequencies of the eigenfunctions in Prob. 3?

5–8 GRAPHING SOLUTIONS

Using (13) sketch or graph a figure (similar to Fig. 291 in Sec. 12.3) of the deflection $u(x, t)$ of a vibrating string (length $L = 1$, ends fixed, $c = 1$) starting with initial velocity 0 and initial deflection (k small, say, $k = 0.01$).

5. $f(x) = k \sin \pi x$
6. $f(x) = k(1 - \cos \pi x)$
7. $f(x) = k \sin 2\pi x$
8. $f(x) = kx(1 - x)$

9–18 NORMAL FORMS

Find the type, transform to normal form, and solve. Show your work in detail.

9. $u_{xx} + 4u_{yy} = 0$
10. $u_{xx} - 16u_{yy} = 0$
11. $u_{xx} + 2u_{xy} + u_{yy} = 0$
12. $u_{xx} - 2u_{xy} + u_{yy} = 0$
13. $u_{xx} + 5u_{xy} + 4u_{yy} = 0$
14. $xu_{xy} - yu_{yy} = 0$
15. $xu_{xx} - yu_{xy} = 0$
16. $u_{xx} + 2u_{xy} + 10u_{yy} = 0$
17. $u_{xx} - 4u_{xy} + 5u_{yy} = 0$
18. $u_{xx} - 6u_{xy} + 9u_{yy} = 0$
19. **Longitudinal Vibrations of an Elastic Bar or Rod.** These vibrations in the direction of the x-axis are modeled by the wave equation $u_{tt} = c^2 u_{xx}$, $c^2 = E/\rho$ (see Tolstov [C9], p. 275). If the rod is fastened at one end, $x = 0$, and free at the other, $x = L$, we have $u(0, t) = 0$ and $u_x(L, t) = 0$. Show that the motion corresponding to initial displacement $u(x, 0) = f(x)$ and initial velocity zero is

$$u = \sum_{n=0}^{\infty} A_n \sin p_n x \cos p_n ct,$$

$$A_n = \frac{2}{L}\int_0^L f(x) \sin p_n x \, dx, \qquad p_n = \frac{(2n+1)\pi}{2L}.$$

20. **Tricomi and Airy equations.**[2] Show that the *Tricomi equation* $yu_{xx} + u_{yy} = 0$ is of mixed type. Obtain the **Airy equation** $G'' - yG = 0$ from the Tricomi equation by separation. (For solutions, see p. 446 of Ref. [GenRef1] listed in App. 1.)

---

[2]Sir GEORGE BIDELL AIRY (1801–1892), English mathematician, known for his work in elasticity. FRANCESCO TRICOMI (1897–1978), Italian mathematician, who worked in integral equations and functional analysis.

12.5 Modeling: Heat Flow from a Body in Space. Heat Equation

After the wave equation (Sec. 12.2) we now derive and discuss the next "big" PDE, the **heat equation**, which governs the temperature u in a body in space. We obtain this model of temperature distribution under the following.

Physical Assumptions

1. The *specific heat* σ and the *density* ρ of the material of the body are constant. No heat is produced or disappears in the body.
2. Experiments show that, in a body, heat flows in the direction of decreasing temperature, and the rate of flow is proportional to the gradient (cf. Sec. 9.7) of the temperature; that is, the velocity $\mathbf{v}$ of the heat flow in the body is of the form

$$\mathbf{v} = -K \operatorname{grad} u \tag{1}$$

   where $u(x, y, z, t)$ is the temperature at a point (x, y, z) and time t.
3. The *thermal conductivity* K is constant, as is the case for homogeneous material and nonextreme temperatures.

Under these assumptions we can model heat flow as follows.

Let T be a region in the body bounded by a surface S with outer unit normal vector $\mathbf{n}$ such that the divergence theorem (Sec. 10.7) applies. Then

$$\mathbf{v} \cdot \mathbf{n}$$

is the component of $\mathbf{v}$ in the direction of $\mathbf{n}$. Hence $|\mathbf{v} \cdot \mathbf{n}\, \Delta A|$ is the amount of heat *leaving* T (if $\mathbf{v} \cdot \mathbf{n} > 0$ at some point P) or *entering* T (if $\mathbf{v} \cdot \mathbf{n} < 0$ at P) per unit time at some point P of S through a small portion ΔS of S of area ΔA. Hence the total amount of heat that flows across S from T is given by the surface integral

$$\iint_S \mathbf{v} \cdot \mathbf{n}\, dA.$$

Note that, so far, this parallels the derivation on fluid flow in Example 1 of Sec. 10.8.

Using Gauss's theorem (Sec. 10.7), we now convert our surface integral into a volume integral over the region T. Because of (1) this gives [use (3) in Sec. 9.8]

$$\begin{aligned} \iint_S \mathbf{v} \cdot \mathbf{n}\, dA &= -K \iint_S (\operatorname{grad} u) \cdot \mathbf{n}\, dA = -K \iiint_T \operatorname{div} (\operatorname{grad} u)\, dx\, dy\, dz \\ &= -K \iiint_T \nabla^2 u\, dx\, dy\, dz. \end{aligned} \tag{2}$$

Here,

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}$$

is the **Laplacian** of u.

On the other hand, the total amount of heat in T is

$$H = \iiint_T \sigma \rho u \, dx \, dy \, dz$$

with σ and ρ as before. Hence the time rate of decrease of H is

$$-\frac{\partial H}{\partial t} = -\iiint_T \sigma \rho \frac{\partial u}{\partial t} \, dx \, dy \, dz.$$

This must be equal to the amount of heat leaving T because no heat is produced or disappears in the body. From (2) we thus obtain

$$-\iiint_T \sigma \rho \frac{\partial u}{\partial t} \, dx \, dy \, dz = -K \iiint_T \nabla^2 u \, dx \, dy \, dz$$

or (divide by $-\sigma\rho$)

$$\iiint_T \left(\frac{\partial u}{\partial t} - c^2 \nabla^2 u \right) dx \, dy \, dz = 0 \qquad c^2 = \frac{K}{\sigma \rho}.$$

Since this holds for any region T in the body, the integrand (if continuous) must be zero everywhere. That is,

$$\textbf{(3)} \qquad \frac{\partial u}{\partial t} = c^2 \nabla^2 u. \qquad c^2 = K/\rho\sigma$$

This is the **heat equation**, the fundamental PDE modeling heat flow. It gives the temperature $u(x, y, z, t)$ in a body of homogeneous material in space. The constant c^2 is the *thermal diffusivity*. K is the *thermal conductivity*, σ the *specific heat*, and ρ the *density* of the material of the body. $\nabla^2 u$ is the Laplacian of u and, with respect to the Cartesian coordinates x, y, z, is

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

The heat equation is also called the **diffusion equation** because it also models chemical diffusion processes of one substance or gas into another.

12.6 Heat Equation: Solution by Fourier Series. Steady Two-Dimensional Heat Problems. Dirichlet Problem

We want to solve the (one-dimensional) heat equation just developed in Sec. 12.5 and give several applications. This is followed much later in this section by an extension of the heat equation to two dimensions.

Fig. 294. Bar under consideration

As an important application of the heat equation, let us first consider the temperature in a long thin metal bar or wire of constant cross section and homogeneous material, which is oriented along the x-axis (Fig. 294) and is perfectly insulated laterally, so that heat flows in the x-direction only. Then besides time, u depends only on x, so that the Laplacian reduces to $u_{xx} = \partial^2 u/\partial x^2$, and the heat equation becomes the **one-dimensional heat equation**

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2}. \tag{1}$$

This PDE seems to differ only very little from the wave equation, which has a term u_{tt} instead of u_t, but we shall see that this will make the solutions of (1) behave quite differently from those of the wave equation.

We shall solve (1) for some important types of boundary and initial conditions. We begin with the case in which the ends $x = 0$ and $x = L$ of the bar are kept at temperature zero, so that we have the **boundary conditions**

$$u(0, t) = 0, \qquad u(L, t) = 0 \qquad \text{for all } t \geqq 0. \tag{2}$$

Furthermore, the initial temperature in the bar at time $t = 0$ is given, say, $f(x)$, so that we have the **initial condition**

$$u(x, 0) = f(x) \qquad [f(x) \text{ given}]. \tag{3}$$

Here we must have $f(0) = 0$ and $f(L) = 0$ because of (2).

We shall determine a solution $u(x, t)$ of (1) satisfying (2) and (3)—one initial condition will be enough, as opposed to two initial conditions for the wave equation. Technically, our method will parallel that for the wave equation in Sec. 12.3: a separation of variables, followed by the use of Fourier series. You may find a step-by-step comparison worthwhile.

Step 1. **Two ODEs from the heat equation (1).** Substitution of a product $u(x, t) = F(x)G(t)$ into (1) gives $F\dot{G} = c^2 F''G$ with $\dot{G} = dG/dt$ and $F'' = d^2F/dx^2$. To separate the variables, we divide by c^2FG, obtaining

$$\frac{\dot{G}}{c^2 G} = \frac{F''}{F}. \tag{4}$$

The left side depends only on t and the right side only on x, so that both sides must equal a constant k (as in Sec. 12.3). You may show that for $k = 0$ or $k > 0$ the only solution $u = FG$ satisfying (2) is $u \equiv 0$. For negative $k = -p^2$ we have from (4)

$$\frac{\dot{G}}{c^2 G} = \frac{F''}{F} = -p^2.$$

Multiplication by the denominators immediately gives the two ODEs

$$F'' + p^2F = 0 \tag{5}$$

and

$$\dot{G} + c^2p^2G = 0. \tag{6}$$

***Step 2.* Satisfying the boundary conditions (2).** We first solve (5). A general solution is

$$F(x) = A \cos px + B \sin px. \tag{7}$$

From the boundary conditions (2) it follows that

$$u(0, t) = F(0)G(t) = 0 \qquad \text{and} \qquad u(L, t) = F(L)G(t) = 0.$$

Since $G \equiv 0$ would give $u \equiv 0$, we require $F(0) = 0$, $F(L) = 0$ and get $F(0) = A = 0$ by (7) and then $F(L) = B \sin pL = 0$, with $B \neq 0$ (to avoid $F \equiv 0$); thus,

$$\sin pL = 0, \qquad \text{hence} \qquad p = \frac{n\pi}{L}, \qquad n = 1, 2, \cdots.$$

Setting $B = 1$, we thus obtain the following solutions of (5) satisfying (2):

$$F_n(x) = \sin \frac{n\pi x}{L}, \qquad n = 1, 2, \cdots.$$

(As in Sec. 12.3, we need not consider *negative* integer values of n.)

All this was literally the same as in Sec. 12.3. From now on it differs since (6) differs from (6) in Sec. 12.3. We now solve (6). For $p = n\pi/L$, as just obtained, (6) becomes

$$\dot{G} + \lambda_n^2 G = 0 \qquad \text{where} \qquad \lambda_n = \frac{cn\pi}{L}.$$

It has the general solution

$$G_n(t) = B_n e^{-\lambda_n^2 t}, \qquad n = 1, 2, \cdots$$

where B_n is a constant. Hence the functions

$$u_n(x, t) = F_n(x)G_n(t) = B_n \sin \frac{n\pi x}{L} e^{-\lambda_n^2 t} \qquad (n = 1, 2, \cdots) \tag{8}$$

are solutions of the heat equation (1), satisfying (2). These are the **eigenfunctions** of the problem, corresponding to the **eigenvalues** $\lambda_n = cn\pi/L$.

***Step 3.* Solution of the entire problem. Fourier series.** So far we have solutions (8) satisfying the boundary conditions (2). To obtain a solution that also satisfies the initial condition (3), we consider a series of these eigenfunctions,

$$u(x, t) = \sum_{n=1}^{\infty} u_n(x, t) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L} e^{-\lambda_n^2 t} \qquad \left(\lambda_n = \frac{cn\pi}{L}\right). \tag{9}$$

From this and (3) we have

$$u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{L} = f(x).$$

Hence for (9) to satisfy (3), the B_n's must be the coefficients of the **Fourier sine series**, as given by (4) in Sec. 11.3; thus

$$B_n = \frac{2}{L}\int_0^L f(x) \sin \frac{n\pi x}{L}\, dx \qquad (n = 1, 2, \cdots) \tag{10}$$

The solution of our problem can be established, assuming that $f(x)$ is piecewise continuous (see Sec. 6.1) on the interval $0 \leqq x \leqq L$ and has one-sided derivatives (see Sec. 11.1) at all interior points of that interval; that is, under these assumptions the series (9) with coefficients (10) is the solution of our physical problem. A proof requires knowledge of uniform convergence and will be given at a later occasion (Probs. 19, 20 in Problem Set 15.5).

Because of the exponential factor, all the terms in (9) approach zero as t approaches infinity. The rate of decay increases with n.

EXAMPLE 1 Sinusoidal Initial Temperature

Find the temperature $u(x, t)$ in a laterally insulated copper bar 80 cm long if the initial temperature is $100 \sin(\pi x/80)$ °C and the ends are kept at 0°C. How long will it take for the maximum temperature in the bar to drop to 50°C? First guess, then calculate. *Physical data for copper:* density 8.92 g/cm^3, specific heat 0.092 cal/(g °C), thermal conductivity 0.95 cal/(cm sec °C).

Solution. The initial condition gives

$$u(x, 0) = \sum_{n=1}^{\infty} B_n \sin \frac{n\pi x}{80} = f(x) = 100 \sin \frac{\pi x}{80}.$$

Hence, by inspection or from (9), we get $B_1 = 100$, $B_2 = B_3 = \cdots = 0$. In (9) we need $\lambda_1^2 = c^2\pi^2/L^2$, where $c^2 = K/(\sigma\rho) = 0.95/(0.092 \cdot 8.92) = 1.158$ [cm^2/sec]. Hence we obtain

$$\lambda_1^2 = 1.158 \cdot 9.870/80^2 = 0.001785 \text{ [sec}^{-1}\text{]}.$$

The solution (9) is

$$u(x, t) = 100 \sin \frac{\pi x}{80} e^{-0.001785t}.$$

Also, $100e^{-0.001785t} = 50$ when $t = (\ln 0.5)/(-0.001785) = 388$ [sec] ≈ 6.5 [min]. Does your guess, or at least its order of magnitude, agree with this result? ∎

EXAMPLE 2 Speed of Decay

Solve the problem in Example 1 when the initial temperature is $100 \sin(3\pi x/80)$ °C and the other data are as before.

Solution. In (9), instead of $n = 1$ we now have $n = 3$, and $\lambda_3^2 = 3^2\lambda_1^2 = 9 \cdot 0.001785 = 0.01607$, so that the solution now is

$$u(x, t) = 100 \sin \frac{3\pi x}{80} e^{-0.01607t}.$$

Hence the maximum temperature drops to 50°C in $t = (\ln 0.5)/(-0.01607) \approx 43$ [sec], which is much faster (9 times as fast as in Example 1; why?).

Had we chosen a bigger n, the decay would have been still faster, and in a sum or series of such terms, each term has its own rate of decay, and terms with large n are practically 0 after a very short time. Our next example is of this type, and the curve in Fig. 295 corresponding to $t = 0.5$ looks almost like a sine curve; that is, it is practically the graph of the first term of the solution. ■

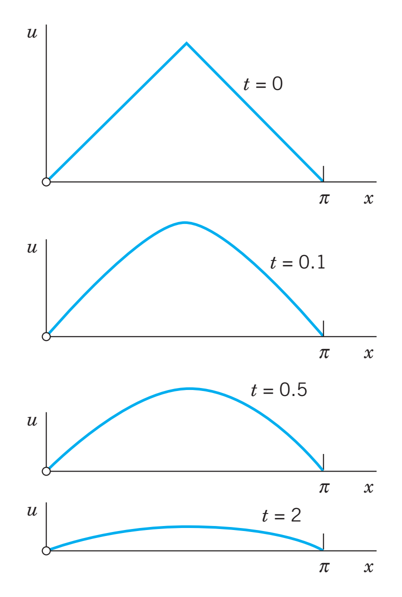


Fig. 295. Example 3. Decrease of temperature with time t for $L = \pi$ and $c = 1$

EXAMPLE 3 **"Triangular" Initial Temperature in a Bar**

Find the temperature in a laterally insulated bar of length L whose ends are kept at temperature 0, assuming that the initial temperature is

$$f(x) = \begin{cases} x & \text{if} \quad 0 < x < L/2, \\ L - x & \text{if} \quad L/2 < x < L. \end{cases}$$

(The uppermost part of Fig. 295 shows this function for the special $L = \pi$.)

Solution. From (10) we get

$$B_n = \frac{2}{L}\left(\int_0^{L/2} x \sin\frac{n\pi x}{L}\,dx + \int_{L/2}^{L} (L - x)\sin\frac{n\pi x}{L}\,dx\right). \tag{10*}$$

Integration gives $B_n = 0$ if n is even,

$$B_n = \frac{4L}{n^2\pi^2} \quad (n = 1, 5, 9, \cdots) \qquad \text{and} \qquad B_n = -\frac{4L}{n^2\pi^2} \quad (n = 3, 7, 11, \cdots).$$

(see also Example 4 in Sec. 11.3 with $k = L/2$). Hence the solution is

$$u(x, t) = \frac{4L}{\pi^2}\left[\sin\frac{\pi x}{L}\exp\left[-\left(\frac{c\pi}{L}\right)^2 t\right] - \frac{1}{9}\sin\frac{3\pi x}{L}\exp\left[-\left(\frac{3c\pi}{L}\right)^2 t\right] + - \cdots\right].$$

Figure 295 shows that the temperature decreases with increasing t, because of the heat loss due to the cooling of the ends.

Compare Fig. 295 and Fig. 291 in Sec. 12.3 and comment. ■

EXAMPLE 4 **Bar with Insulated Ends. Eigenvalue 0**

Find a solution formula of (1), (3) with (2) replaced by the condition that both ends of the bar are insulated.

Solution. Physical experiments show that the rate of heat flow is proportional to the gradient of the temperature. Hence if the ends $x = 0$ and $x = L$ of the bar are insulated, so that no heat can flow through the ends, we have grad $u = u_x = \partial u/\partial x$ and the boundary conditions

$$(2^*) \qquad u_x(0, t) = 0, \qquad u_x(L, t) = 0 \qquad \text{for all } t.$$

Since $u(x, t) = F(x)G(t)$, this gives $u_x(0, t) = F'(0)G(t) = 0$ and $u_x(L, t) = F'(L)G(t) = 0$. Differentiating (7), we have $F'(x) = -Ap \sin px + Bp \cos px$, so that

$$F'(0) = Bp = 0 \qquad \text{and then} \qquad F'(L) = -Ap \sin pL = 0.$$

The second of these conditions gives $p = p_n = n\pi/L$, $(n = 0, 1, 2, \cdots)$. From this and (7) with $A = 1$ and $B = 0$ we get $F_n(x) = \cos(n\pi x/L)$, $(n = 0, 1, 2, \cdots)$. With G_n as before, this yields the eigenfunctions

$$(11) \qquad u_n(x, t) = F_n(x)G_n(t) = A_n \cos\frac{n\pi x}{L} e^{-\lambda_n^2 t} \qquad (n = 0, 1, \cdots)$$

corresponding to the eigenvalues $\lambda_n = cn\pi/L$. The latter are as before, but we now have the additional eigenvalue $\lambda_0 = 0$ and eigenfunction $u_0 = \text{const}$, which is the solution of the problem if the initial temperature $f(x)$ is constant. This shows the remarkable fact that ***a separation constant can very well be zero, and zero can be an eigenvalue.***

Furthermore, whereas (8) gave a Fourier sine series, we now get from (11) a Fourier cosine series

$$\textbf{(12)} \qquad u(x, t) = \sum_{n=0}^{\infty} u_n(x, t) = \sum_{n=0}^{\infty} A_n \cos\frac{n\pi x}{L} e^{-\lambda_n^2 t} \qquad \left(\lambda_n = \frac{cn\pi}{L}\right).$$

Its coefficients result from the initial condition (3),

$$u(x, 0) = \sum_{n=0}^{\infty} A_n \cos\frac{n\pi x}{L} = f(x),$$

in the form (2), Sec. 11.3, that is,

$$(13) \qquad A_0 = \frac{1}{L}\int_0^L f(x)\,dx, \qquad A_n = \frac{2}{L}\int_0^L f(x)\cos\frac{n\pi x}{L}\,dx, \qquad n = 1, 2, \cdots.$$

■

EXAMPLE 5 **"Triangular" Initial Temperature in a Bar with Insulated Ends**

Find the temperature in the bar in Example 3, assuming that the ends are insulated (instead of being kept at temperature 0).

Solution. For the triangular initial temperature, (13) gives $A_0 = L/4$ and (see also Example 4 in Sec. 11.3 with $k = L/2$)

$$A_n = \frac{2}{L}\left[\int_0^{L/2} x \cos\frac{n\pi x}{L}\,dx + \int_{L/2}^{L} (L - x)\cos\frac{n\pi x}{L}\,dx\right] = \frac{2L}{n^2\pi^2}\left(2\cos\frac{n\pi}{2} - \cos n\pi - 1\right).$$

Hence the solution (12) is

$$u(x, t) = \frac{L}{4} - \frac{8L}{\pi^2}\left\{\frac{1}{2^2}\cos\frac{2\pi x}{L}\exp\left[-\left(\frac{2c\pi}{L}\right)^2 t\right] + \frac{1}{6^2}\cos\frac{6\pi x}{L}\exp\left[-\left(\frac{6c\pi}{L}\right)^2 t\right] + \cdots\right\}.$$

We see that the terms decrease with increasing t, and $u \to L/4$ as $t \to \infty$; this is the mean value of the initial temperature. This is plausible because no heat can escape from this totally insulated bar. In contrast, the cooling of the ends in Example 3 led to heat loss and $u \to 0$, the temperature at which the ends were kept. ■

Steady Two-Dimensional Heat Problems. Laplace's Equation

We shall now extend our discussion from one to two space dimensions and consider the two-dimensional heat equation

$$\frac{\partial u}{\partial t} = c^2 \nabla^2 u = c^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} \right)$$

for **steady** (that is, *time-independent*) problems. Then $\partial u / \partial t = 0$ and the heat equation reduces to **Laplace's equation**

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \qquad \textbf{(14)}$$

(which has already occurred in Sec. 10.8 and will be considered further in Secs. 12.8–12.11). A heat problem then consists of this PDE to be considered in some region R of the xy-plane and a given boundary condition on the boundary curve C of R. This is a **boundary value problem (BVP)**. One calls it:

First BVP or **Dirichlet Problem** if u is prescribed on C ("**Dirichlet boundary condition**")

Second BVP or **Neumann Problem** if the normal derivative $u_n = \partial u / \partial n$ is prescribed on C ("**Neumann boundary condition**")

Third BVP, **Mixed BVP**, or **Robin Problem** if u is prescribed on a portion of C and u_n on the rest of C ("**Mixed boundary condition**").

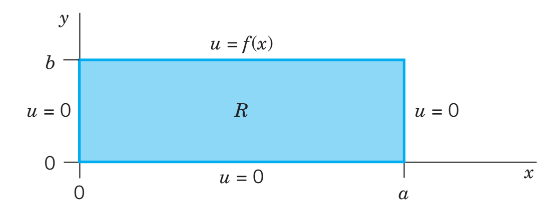


Fig. 296. Rectangle R and given boundary values

Dirichlet Problem in a Rectangle R (Fig. 296). We consider a Dirichlet problem for Laplace's equation (14) in a rectangle R, assuming that the temperature $u(x, y)$ equals a given function $f(x)$ on the upper side and 0 on the other three sides of the rectangle.

We solve this problem by separating variables. Substituting $u(x, y) = F(x)G(y)$ into (14) written as $u_{xx} = -u_{yy}$, dividing by FG, and equating both sides to a negative constant, we obtain

$$\frac{1}{F} \cdot \frac{d^2F}{dx^2} = -\frac{1}{G} \cdot \frac{d^2G}{dy^2} = -k.$$

From this we get

$$\frac{d^2F}{dx^2} + kF = 0,$$

and the left and right boundary conditions imply

$$F(0) = 0, \qquad \text{and} \qquad F(a) = 0.$$

This gives $k = (n\pi/a)^2$ and corresponding nonzero solutions

(15) $$F(x) = F_n(x) = \sin \frac{n\pi}{a} x, \qquad n = 1, 2, \cdots.$$

The ODE for G with $k = (n\pi/a)^2$ then becomes

$$\frac{d^2G}{dy^2} - \left(\frac{n\pi}{a}\right)^2 G = 0.$$

Solutions are

$$G(y) = G_n(y) = A_n e^{n\pi y/a} + B_n e^{-n\pi y/a}.$$

Now the boundary condition $u = 0$ on the lower side of R implies that $G_n(0) = 0$; that is, $G_n(0) = A_n + B_n = 0$ or $B_n = -A_n$. This gives

$$G_n(y) = A_n(e^{n\pi y/a} - e^{-n\pi y/a}) = 2A_n \sinh \frac{n\pi y}{a}.$$

From this and (15), writing $2A_n = A_n^*$, we obtain as the **eigenfunctions** of our problem

(16) $$u_n(x, y) = F_n(x)G_n(y) = A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi y}{a}.$$

These solutions satisfy the boundary condition $u = 0$ on the left, right, and lower sides.

To get a solution also satisfying the boundary condition $u(x, b) = f(x)$ on the upper side, we consider the infinite series

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x, y).$$

From this and (16) with $y = b$ we obtain

$$u(x, b) = f(x) = \sum_{n=1}^{\infty} A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi b}{a}.$$

We can write this in the form

$$u(x, b) = \sum_{n=1}^{\infty} \left(A_n^* \sinh \frac{n\pi b}{a}\right) \sin \frac{n\pi x}{a}.$$

This shows that the expressions in the parentheses must be the Fourier coefficients b_n of $f(x)$; that is, by (4) in Sec. 11.3,

$$b_n = A_n^* \sinh \frac{n\pi b}{a} = \frac{2}{a} \int_0^a f(x) \sin \frac{n\pi x}{a}\, dx.$$

From this and (16) we see that the solution of our problem is

$$u(x, y) = \sum_{n=1}^{\infty} u_n(x, y) = \sum_{n=1}^{\infty} A_n^* \sin \frac{n\pi x}{a} \sinh \frac{n\pi y}{a} \tag{17}$$

where

$$A_n^* = \frac{2}{a \sinh (n\pi b/a)} \int_0^a f(x) \sin \frac{n\pi x}{a}\, dx. \tag{18}$$

We have obtained this solution formally, neither considering convergence nor showing that the series for u, u_{xx}, and u_{yy} have the right sums. This can be proved if one assumes that f and f' are continuous and f'' is piecewise continuous on the interval $0 \leqq x \leqq a$. The proof is somewhat involved and relies on uniform convergence. It can be found in [C4] listed in App. 1.

Unifying Power of Methods. Electrostatics, Elasticity

The Laplace equation (14) also governs the electrostatic potential of electrical charges in any region that is free of these charges. Thus our steady-state heat problem can also be interpreted as an electrostatic potential problem. Then (17), (18) is the potential in the rectangle R when the upper side of R is at potential $f(x)$ and the other three sides are grounded.

Actually, in the steady-state case, the two-dimensional wave equation (to be considered in Secs. 12.8, 12.9) also reduces to (14). Then (17), (18) is the displacement of a rectangular elastic membrane (rubber sheet, drumhead) that is fixed along its boundary, with three sides lying in the xy-plane and the fourth side given the displacement $f(x)$.

This is another impressive demonstration of the ***unifying power*** of mathematics. It illustrates that ***entirely different physical systems may have the same mathematical model*** and can thus be treated by the same mathematical methods.

PROBLEM SET 12.6

1. **Decay.** How does the rate of decay of (8) with fixed n depend on the specific heat, the density, and the thermal conductivity of the material?

2. **Decay.** If the first eigenfunction (8) of the bar decreases to half its value within 20 sec, what is the value of the diffusivity?

3. **Eigenfunctions.** Sketch or graph and compare the first three eigenfunctions (8) with $B_n = 1, c = 1$, and $L = \pi$ for $t = 0, 0.1, 0.2, \cdots, 1.0$.

4. **WRITING PROJECT. Wave and Heat Equations.** Compare these PDEs with respect to general behavior of eigenfunctions and kind of boundary and initial

conditions. State the difference between Fig. 291 in Sec. 12.3 and Fig. 295.

5–7 LATERALLY INSULATED BAR

Find the temperature $u(x, t)$ in a bar of silver of length 10 cm and constant cross section of area 1 cm^2 (density 10.6 g/cm^3, thermal conductivity 1.04 cal/(cm sec °C), specific heat 0.056 cal/(g °C) that is perfectly insulated laterally, with ends kept at temperature 0°C and initial temperature $f(x)$ °C, where

5. $f(x) = \sin 0.1\pi x$
6. $f(x) = 4 - 0.8|x - 5|$
7. $f(x) = x(10 - x)$
8. **Arbitrary temperatures at ends.** If the ends $x = 0$ and $x = L$ of the bar in the text are kept at constant temperatures U_1 and U_2, respectively, what is the temperature $u_1(x)$ in the bar after a long time (theoretically, as $t \to \infty$)? First guess, then calculate.
9. In Prob. 8 find the temperature at any time.
10. **Change of end temperatures.** Assume that the ends of the bar in Probs. 5–7 have been kept at 100°C for a long time. Then at some instant, call it $t = 0$, the temperature at $x = L$ is suddenly changed to 0°C and kept at 0°C, whereas the temperature at $x = 0$ is kept at 100°C. Find the temperature in the middle of the bar at $t = 1, 2, 3, 10, 50$ sec. First guess, then calculate.

BAR UNDER ADIABATIC CONDITIONS

"Adiabatic" means no heat exchange with the neighborhood, because the bar is completely insulated, also at the ends. *Physical Information:* The heat flux at the ends is proportional to the value of $\partial u/\partial x$ there.

11. Show that for the completely insulated bar, $u_x(0, t) = 0$, $u_x(L, t) = 0$, $u(x, t) = f(x)$ and separation of variables gives the following solution, with A_n given by (2) in Sec. 11.3.

$$u(x, t) = A_0 + \sum_{n=1}^{\infty} A_n \cos\frac{n\pi x}{L}\, e^{-(cn\pi/L)^2 t}$$

12–15 Find the temperature in Prob. 11 with $L = \pi$, $c = 1$, and

12. $f(x) = x$
13. $f(x) = 1$
14. $f(x) = \cos 2x$
15. $f(x) = 1 - x/\pi$
16. **A bar with heat generation** of constant rate $H\ (>0)$ is modeled by $u_t = c^2 u_{xx} + H$. Solve this problem if $L = \pi$ and the ends of the bar are kept at 0°C. *Hint.* Set $u = v - Hx(x - \pi)/(2c^2)$.
17. **Heat flux.** The *heat flux* of a solution $u(x, t)$ across $x = 0$ is defined by $\phi(t) = -Ku_x(0, t)$. Find $\phi(t)$ for the solution (9). Explain the name. Is it physically understandable that ϕ goes to 0 as $t \to \infty$?

18–25 TWO-DIMENSIONAL PROBLEMS

18. **Laplace equation.** Find the potential in the rectangle $0 \leqq x \leqq 20$, $0 \leqq y \leqq 40$ whose upper side is kept at potential 110 V and whose other sides are grounded.
19. Find the potential in the square $0 \leqq x \leqq 2$, $0 \leqq y \leqq 2$ if the upper side is kept at the potential $1000 \sin \frac{1}{2}\pi x$ and the other sides are grounded.
20. **CAS PROJECT. Isotherms.** Find the steady-state solutions (temperatures) in the square plate in Fig. 297 with $a = 2$ satisfying the following boundary conditions. Graph isotherms.
    (a) $u = 80 \sin \pi x$ on the upper side, 0 on the others.
    (b) $u = 0$ on the vertical sides, assuming that the other sides are perfectly insulated.
    (c) Boundary conditions of your choice (such that the solution is not identically zero).

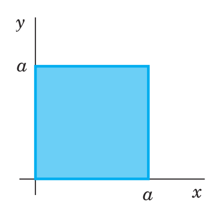


Fig. 297. Square plate

21. **Heat flow in a plate.** The faces of the thin square plate in Fig. 297 with side $a = 24$ are perfectly insulated. The upper side is kept at 25°C and the other sides are kept at 0°C. Find the steady-state temperature $u(x, y)$ in the plate.
22. Find the steady-state temperature in the plate in Prob. 21 if the lower side is kept at U_0°C, the upper side at U_1°C, and the other sides are kept at 0°C. *Hint:* Split into two problems in which the boundary temperature is 0 on three sides for each problem.
23. **Mixed boundary value problem.** Find the steady-state temperature in the plate in Prob. 21 with the upper and lower sides perfectly insulated, the left side kept at 0°C, and the right side kept at $f(y)$°C.
24. **Radiation.** Find steady-state temperatures in the rectangle in Fig. 296 with the upper and left sides perfectly insulated and the right side radiating into a medium at 0°C according to $u_x(a, y) + hu(a, y) = 0$, $h > 0$ constant. (You will get many solutions since no condition on the lower side is given.)
25. Find formulas similar to (17), (18) for the temperature in the rectangle R of the text when the lower side of R is kept at temperature $f(x)$ and the other sides are kept at 0°C.

12.7 Heat Equation: Modeling Very Long Bars. Solution by Fourier Integrals and Transforms

Our discussion of the heat equation

$$\frac{\partial u}{\partial t} = c^2 \frac{\partial^2 u}{\partial x^2} \tag{1}$$

in the last section extends to bars of infinite length, which are good models of very long bars or wires (such as a wire of length, say, 300 ft). Then the role of Fourier series in the solution process will be taken by **Fourier integrals** (Sec. 11.7).

Let us illustrate the method by solving (1) for a bar that extends to infinity on both sides (and is laterally insulated as before). Then we do not have boundary conditions, but only the **initial condition**

$$u(x, 0) = f(x) \qquad (-\infty < x < \infty) \tag{2}$$

where $f(x)$ is the given initial temperature of the bar.

To solve this problem, we start as in the last section, substituting $u(x, t) = F(x)G(t)$ into (1). This gives the two ODEs

$$F'' + p^2 F = 0 \qquad \text{[see (5), Sec. 12.6]} \tag{3}$$

and

$$\dot{G} + c^2 p^2 G = 0 \qquad \text{[see (6), Sec. 12.6].} \tag{4}$$

Solutions are

$$F(x) = A \cos px + B \sin px \qquad \text{and} \qquad G(t) = e^{-c^2p^2t},$$

respectively, where A and B are any constants. Hence a solution of (1) is

$$u(x, t; p) = FG = (A \cos px + B \sin px)\, e^{-c^2p^2t}. \tag{5}$$

Here we had to choose the separation constant k negative, $k = -p^2$, because positive values of k would lead to an increasing exponential function in (5), which has no physical meaning.

Use of Fourier Integrals

Any series of functions (5), found in the usual manner by taking p as multiples of a fixed number, would lead to a function that is periodic in x when $t = 0$. However, since $f(x)$

in (2) is not assumed to be periodic, it is natural to use **Fourier integrals** instead of Fourier series. Also, A and B in (5) are arbitrary and we may regard them as functions of p, writing $A = A(p)$ and $B = B(P)$. Now, since the heat equation (1) is linear and homogeneous, the function

$$u(x, t) = \int_0^\infty u(x, t; p)\, dp = \int_0^\infty [A(p) \cos px + B(p) \sin px]\, e^{-c^2p^2t}\, dp \tag{6}$$

is then a solution of (1), provided this integral exists and can be differentiated twice with respect to x and once with respect to t.

Determination of $A(p)$ and $B(p)$ from the Initial Condition. From (6) and (2) we get

$$u(x, 0) = \int_0^\infty [A(p) \cos px + B(p) \sin px]\, dp = f(x). \tag{7}$$

This gives $A(p)$ and $B(p)$ in terms of $f(x)$; indeed, from (4) in Sec. 11.7 we have

$$A(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v) \cos pv\, dv, \qquad B(p) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v) \sin pv\, dv. \tag{8}$$

According to (1*), Sec. 11.9, our Fourier integral (7) with these $A(p)$ and $B(p)$ can be written

$$u(x, 0) = \frac{1}{\pi}\int_0^\infty \left[\int_{-\infty}^{\infty} f(v) \cos(px - pv)\, dv\right] dp.$$

Similarly, (6) in this section becomes

$$u(x, t) = \frac{1}{\pi}\int_0^\infty \left[\int_{-\infty}^{\infty} f(v) \cos(px - pv)\, e^{-c^2p^2t}\, dv\right] dp.$$

Assuming that we may reverse the order of integration, we obtain

$$u(x, t) = \frac{1}{\pi}\int_{-\infty}^{\infty} f(v)\left[\int_0^\infty e^{-c^2p^2t} \cos(px - pv)\, dp\right] dv. \tag{9}$$

Then we can evaluate the inner integral by using the formula

$$\int_0^\infty e^{-s^2} \cos 2bs\, ds = \frac{\sqrt{\pi}}{2}\, e^{-b^2}. \tag{10}$$

[A derivation of (10) is given in Problem Set 16.4 (Team Project 24).] This takes the form of our inner integral if we choose $p = s/(c\sqrt{t})$ as a new variable of integration and set

$$b = \frac{x - v}{2c\sqrt{t}}.$$

Then $2bs = (x - v)p$ and $ds = c\sqrt{t}\,dp$, so that (10) becomes

$$\int_0^\infty e^{-c^2p^2t} \cos(px - pv)\, dp = \frac{\sqrt{\pi}}{2c\sqrt{t}} \exp\left\{-\frac{(x-v)^2}{4c^2t}\right\}.$$

By inserting this result into (9) we obtain the representation

(11) $$u(x, t) = \frac{1}{2c\sqrt{\pi t}} \int_{-\infty}^{\infty} f(v) \exp\left\{-\frac{(x-v)^2}{4c^2t}\right\} dv.$$

Taking $z = (v - x)/(2c\sqrt{t})$ as a variable of integration, we get the alternative form

(12) $$u(x, t) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\infty} f(x + 2cz\sqrt{t})\, e^{-z^2}\, dz.$$

If $f(x)$ is bounded for all values of x and integrable in every finite interval, it can be shown (see Ref. [C10]) that the function (11) or (12) satisfies (1) and (2). Hence this function is the required solution in the present case.

EXAMPLE 1 Temperature in an Infinite Bar

Find the temperature in the infinite bar if the initial temperature is (Fig. 298)

$$f(x) = \begin{cases} U_0 = \text{const} & \text{if} \quad |x| < 1, \\ 0 & \text{if} \quad |x| > 1. \end{cases}$$

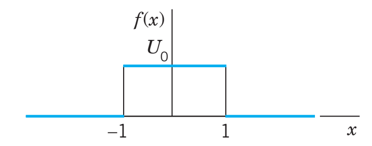


Fig. 298. Initial temperature in Example 1

Solution. From (11) we have

$$u(x, t) = \frac{U_0}{2c\sqrt{\pi t}} \int_{-1}^{1} \exp\left\{-\frac{(x-v)^2}{4c^2t}\right\} dv.$$

If we introduce the above variable of integration z, then the integration over v from -1 to 1 corresponds to the integration over z from $(-1 - x)/(2c\sqrt{t})$ to $(1 - x)/(2c\sqrt{t})$, and

(13) $$u(x, t) = \frac{U_0}{\sqrt{\pi}} \int_{-(1+x)/(2c\sqrt{t})}^{(1-x)/(2c\sqrt{t})} e^{-z^2} dz \qquad (t > 0).$$

We mention that this integral is not an elementary function, but can be expressed in terms of the error function, whose values have been tabulated. (Table A4 in App. 5 contains a few values; larger tables are listed in Ref. [GenRef1] in App. 1. See also CAS Project 1, p. 574.) Figure 299 shows $u(x, t)$ for $U_0 = 100°C$, $c^2 = 1\ \text{cm}^2/\text{sec}$, and several values of t. ■

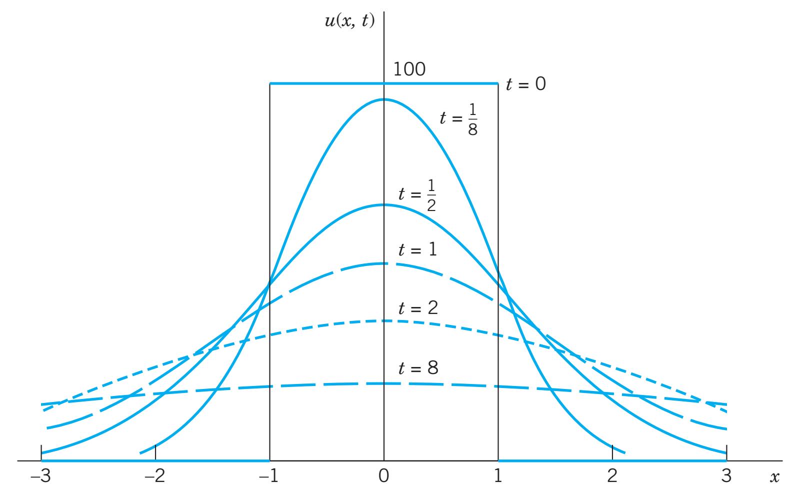


Fig. 299. Solution $u(x, t)$ in Example 1 for $U_0 = 100°\text{C}$, $c^2 = 1 \text{ cm}^2/\text{sec}$, and several values of t

Use of Fourier Transforms

The Fourier transform is closely related to the Fourier integral, from which we obtained the transform in Sec. 11.9. And the transition to the Fourier cosine and sine transform in Sec. 11.8 was even simpler. (You may perhaps wish to review this before going on.) Hence it should not surprise you that we can use these transforms for solving our present or similar problems. The Fourier transform applies to problems concerning the entire axis, and the Fourier cosine and sine transforms to problems involving the positive half-axis. Let us explain these transform methods by typical applications that fit our present discussion.

EXAMPLE 2 Temperature in the Infinite Bar in Example 1

Solve Example 1 using the Fourier transform.

Solution. The problem consists of the heat equation (1) and the initial condition (2), which in this example is

$$f(x) = U_0 = \text{const} \quad \text{if } |x| < 1 \qquad \text{and 0 otherwise.}$$

Our strategy is to take the Fourier transform with respect to x and then to solve the resulting ***ordinary*** DE in t. The details are as follows.

Let $\hat{u} = \mathcal{F}(u)$ denote the Fourier transform of u, ***regarded as a function of x***. From (10) in Sec. 11.9 we see that the heat equation (1) gives

$$\mathcal{F}(u_t) = c^2\mathcal{F}(u_{xx}) = c^2(-w^2)\mathcal{F}(u) = -c^2w^2\hat{u}.$$

On the left, assuming that we may interchange the order of differentiation and integration, we have

$$\mathcal{F}(u_t) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} u_t e^{-iwx}\,dx = \frac{1}{\sqrt{2\pi}}\frac{\partial}{\partial t}\int_{-\infty}^{\infty} u e^{-iwx}\,dx = \frac{\partial \hat{u}}{\partial t}.$$

Thus

$$\frac{\partial \hat{u}}{\partial t} = -c^2w^2\hat{u}.$$

Since this equation involves only a derivative with respect to t but none with respect to w, this is a first-order ***ordinary DE***, with t as the independent variable and w as a parameter. By separating variables (Sec. 1.3) we get the general solution

$$\hat{u}(w, t) = C(w)e^{-c^2w^2t}$$

with the arbitrary "constant" $C(w)$ depending on the parameter w. The initial condition (2) yields the relationship $\hat{u}(w, 0) = C(w) = \hat{f}(w) = \mathscr{F}(f)$. Our intermediate result is

$$\hat{u}(w, t) = \hat{f}(w)e^{-c^2w^2t}.$$

The inversion formula (7), Sec. 11.9, now gives the solution

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(w)\, e^{-c^2w^2t}\, e^{iwx}\, dw. \tag{14}$$

In this solution we may insert the Fourier transform

$$\hat{f}(w) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} f(v)e^{ivw} dv.$$

Assuming that we may invert the order of integration, we then obtain

$$u(x, t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} f(v) \left[\int_{-\infty}^{\infty} e^{-c^2w^2t}\, e^{i(wx - wv)} dw \right] dv.$$

By the Euler formula (3). Sec. 11.9, the integrand of the inner integral equals

$$e^{-c^2w^2t} \cos(wx - wv) + ie^{-c^2w^2t} \sin(wx - wv).$$

We see that its imaginary part is an odd function of w, so that its integral is 0. (More precisely, this is the principal part of the integral; see Sec. 16.4.) The real part is an even function of w, so that its integral from $-\infty$ to ∞ equals twice the integral from 0 to ∞:

$$u(x, t) = \frac{1}{\pi} \int_{-\infty}^{\infty} f(v) \left[\int_{0}^{\infty} e^{-c^2w^2t} \cos(wx - wv)\, dw \right] dv.$$

This agrees with (9) (with $p = w$) and leads to the further formulas (11) and (13). ■

EXAMPLE 3 Solution in Example 1 by the Method of Convolution

Solve the heat problem in Example 1 by the method of convolution.

Solution. The beginning is as in Example 2 and leads to (14), that is,

$$u(x, t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \hat{f}(w)e^{-c^2w^2t}e^{iwx}\, dw. \tag{15}$$

Now comes the crucial idea. We recognize that this is of the form (13) in Sec. 11.9, that is,

$$u(x, t) = (f * g)(x) = \int_{-\infty}^{\infty} \hat{f}(w)\hat{g}(w)e^{iwx}\, dw \tag{16}$$

where

$$\hat{g}(w) = \frac{1}{\sqrt{2\pi}} e^{-c^2w^2t}. \tag{17}$$

Since, by the definition of convolution [(11), Sec. 11.9],

$$(f * g)(x) = \int_{-\infty}^{\infty} f(p)g(x - p)\, dp, \tag{18}$$

as our next and last step we must determine the inverse Fourier transform g of $\hat{g}$. For this we can use formula 9 in Table III of Sec. 11.10,

$$\mathscr{F}(e^{-ax^2}) = \frac{1}{\sqrt{2a}} e^{-w^2/(4a)}$$

with a suitable a. With $c^2 t = 1/(4a)$ or $a = 1/(4c^2 t)$, using (17) we obtain

$$\mathscr{F}(e^{-x^2/(4c^2t)}) = \sqrt{2c^2 t}\, e^{-c^2 w^2 t} = \sqrt{2c^2 t}\, \sqrt{2\pi}\, \hat{g}(w).$$

Hence $\hat{g}$ has the inverse

$$\frac{1}{\sqrt{2c^2 t}\, \sqrt{2\pi}} e^{-x^2/(4c^2t)}.$$

Replacing x with $x - p$ and substituting this into (18) we finally have

$$u(x, t) = (f * g)(x) = \frac{1}{2c\sqrt{\pi t}} \int_{-\infty}^{\infty} f(p) \exp\left\{-\frac{(x-p)^2}{4c^2 t}\right\} dp. \tag{19}$$

This solution formula of our problem agrees with (11). We wrote $(f * g)(x)$, without indicating the parameter t with respect to which we did not integrate. ■

EXAMPLE 4 Fourier Sine Transform Applied to the Heat Equation

If a laterally insulated bar extends from $x = 0$ to infinity, we can use the Fourier sine transform. We let the initial temperature be $u(x, 0) = f(x)$ and impose the boundary condition $u(0, t) = 0$. Then from the heat equation and (9b) in Sec. 11.8, since $f(0) = u(0, 0) = 0$, we obtain

$$\mathscr{F}_s(u_t) = \frac{\partial \hat{u}_s}{\partial t} = c^2 \mathscr{F}_s(u_{xx}) = -c^2 w^2 \mathscr{F}_s(u) = -c^2 w^2 \hat{u}_s(w, t).$$

This is a first-order ODE $\partial \hat{u}_s/\partial t + c^2 w^2 \hat{u}_s = 0$. Its solution is

$$\hat{u}_s(w, t) = C(w) e^{-c^2 w^2 t}.$$

From the initial condition $u(x, 0) = f(x)$ we have $\hat{u}_s(w, 0) = \hat{f}_s(w) = C(w)$. Hence

$$\hat{u}_s(w, t) = \hat{f}_s(w) e^{-c^2 w^2 t}.$$

Taking the inverse Fourier sine transform and substituting

$$\hat{f}_s(w) = \sqrt{\frac{2}{\pi}} \int_0^{\infty} f(p) \sin wp\, dp$$

on the right, we obtain the solution formula

$$u(x, t) = \frac{2}{\pi} \int_0^{\infty} \int_0^{\infty} f(p) \sin wp\, e^{-c^2 w^2 t} \sin wx\, dp\, dw. \tag{20}$$

Figure 300 shows (20) with $c = 1$ for $f(x) = 1$ if $0 \leqq x \leqq 1$ and 0 otherwise, graphed over the xt-plane for $0 \leqq x \leqq 2$, $0.01 \leqq t \leqq 1.5$. Note that the curves of $u(x, t)$ for constant t resemble those in Fig. 299. ■

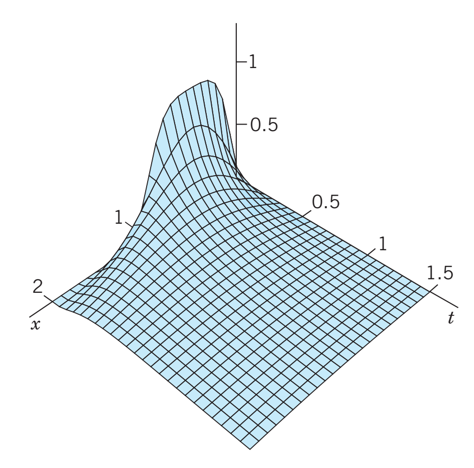


Fig. 300. Solution (20) in Example 4

PROBLEM SET 12.7

1. CAS PROJECT. Heat Flow. (a) Graph the basic Fig. 299.

(b) In (a) apply animation to "see" the heat flow in terms of the decrease of temperature.

(c) Graph $u(x, t)$ with $c = 1$ as a surface over a rectangle of the form $-a < x < a, \quad 0 < y < b$.

2–8 SOLUTION IN INTEGRAL FORM

Using (6), obtain the solution of (1) in integral form satisfying the initial condition $u(x, 0) = f(x)$, where

2. $f(x) = 1$ if $|x| < a$ and 0 otherwise

3. $f(x) = 1/(1 + x^2)$.
Hint. Use (15) in Sec. 11.7.

4. $f(x) = e^{-|x|}$

5. $f(x) = |x|$ if $|x| < 1$ and 0 otherwise

6. $f(x) = x$ if $|x| < 1$ and 0 otherwise

7. $f(x) = (\sin x)/x$.
Hint. Use Prob. 4 in Sec. 11.7.

8. Verify that u in the solution of Prob. 7 satisfies the initial condition.

9–12 CAS PROJECT. Error Function.

$$\text{erf}\, x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-w^2}\, dw \tag{21}$$

This function is important in applied mathematics and physics (probability theory and statistics, thermodynamics, etc.) and fits our present discussion. Regarding it as a typical case of a special function defined by an integral that cannot be evaluated as in elementary calculus, do the following.

9. Graph the **bell-shaped curve** [the curve of the integrand in (21)]. Show that erf x is odd. Show that

$$\int_a^b e^{-w^2}\, dw = \frac{\sqrt{\pi}}{2}(\text{erf}\, b - \text{erf}\, a).$$

$$\int_{-b}^b e^{-w^2}\, dw = \sqrt{\pi}\, \text{erf}\, b.$$

10. Obtain the Maclaurin series of erf x from that of the integrand. Use that series to compute a table of erf x for $x = 0(0.01)3$ (meaning $x = 0, 0.01, 0.02, \cdots, 3$).

11. Obtain the values required in Prob. 10 by an integration command of your CAS. Compare accuracy.

12. It can be shown that erf $(\infty) = 1$. Confirm this experimentally by computing erf x for large x.

13. Let $f(x) = 1$ when $x > 0$ and 0 when $x < 0$. Using erf $(\infty) = 1$, show that (12) then gives

$$u(x, t) = \frac{1}{\sqrt{\pi}} \int_{-x/(2c\sqrt{t})}^{\infty} e^{-x^2} dz$$
$$= \frac{1}{2} - \frac{1}{2}\, \text{erf}\left(-\frac{x}{2c\sqrt{t}}\right) \qquad (t > 0).$$

14. Express the temperature (13) in terms of the error function.

15. Show that $\Phi(x) = \dfrac{1}{\sqrt{2\pi}} \displaystyle\int_{-\infty}^{x} e^{-s^2/2}\, ds$

$$= \frac{1}{2} + \frac{1}{2}\, \text{erf}\left(\frac{x}{\sqrt{2}}\right).$$

Here, the integral is the definition of the "distribution function of the normal probability distribution" to be discussed in Sec. 24.8.

12.8 Modeling: Membrane, Two-Dimensional Wave Equation

Since the modeling here will be similar to that of Sec. 12.2, you may want to take another look at Sec. 12.2.

The vibrating string in Sec. 12.2 is a basic one-dimensional vibrational problem. Equally important is its two-dimensional analog, namely, the motion of an elastic membrane, such as a drumhead, that is stretched and then fixed along its edge. Indeed, setting up the model will proceed almost as in Sec. 12.2.

Physical Assumptions

1. The mass of the membrane per unit area is constant ("homogeneous membrane"). The membrane is perfectly flexible and offers no resistance to bending.
2. The membrane is stretched and then fixed along its entire boundary in the xy-plane. The tension per unit length T caused by stretching the membrane is the same at all points and in all directions and does not change during the motion.
3. The deflection $u(x, y, t)$ of the membrane during the motion is small compared to the size of the membrane, and all angles of inclination are small.

Although these assumptions cannot be realized exactly, they hold relatively accurately for small transverse vibrations of a thin elastic membrane, so that we shall obtain a good model, for instance, of a drumhead.

Derivation of the PDE of the Model ("Two-Dimensional Wave Equation") from Forces. As in Sec. 12.2 the model will consist of a PDE and additional conditions. The PDE will be obtained by the same method as in Sec. 12.2, namely, by considering the forces acting on a small portion of the physical system, the membrane in Fig. 301 on the next page, as it is moving up and down.

Since the deflections of the membrane and the angles of inclination are small, the sides of the portion are approximately equal to Δx and Δy. The tension T is the force per unit length. Hence the forces acting on the sides of the portion are approximately $T\,\Delta x$ and $T\,\Delta y$. Since the membrane is perfectly flexible, these forces are tangent to the moving membrane at every instant.

Horizontal Components of the Forces. We first consider the horizontal components of the forces. These components are obtained by multiplying the forces by the cosines of the angles of inclination. Since these angles are small, their cosines are close to 1. Hence the horizontal components of the forces at opposite sides are approximately equal. Therefore, the motion of the particles of the membrane in a horizontal direction will be negligibly small. From this we conclude that we may regard the motion of the membrane as transversal; that is, each particle moves vertically.

Vertical Components of the Forces. These components along the right side and the left side are (Fig. 301), respectively,

$$T\,\Delta y \sin\beta \qquad \text{and} \qquad -T\,\Delta y \sin\alpha.$$

Here α and β are the values of the angle of inclination (which varies slightly along the edges) in the middle of the edges, and the minus sign appears because the force on the

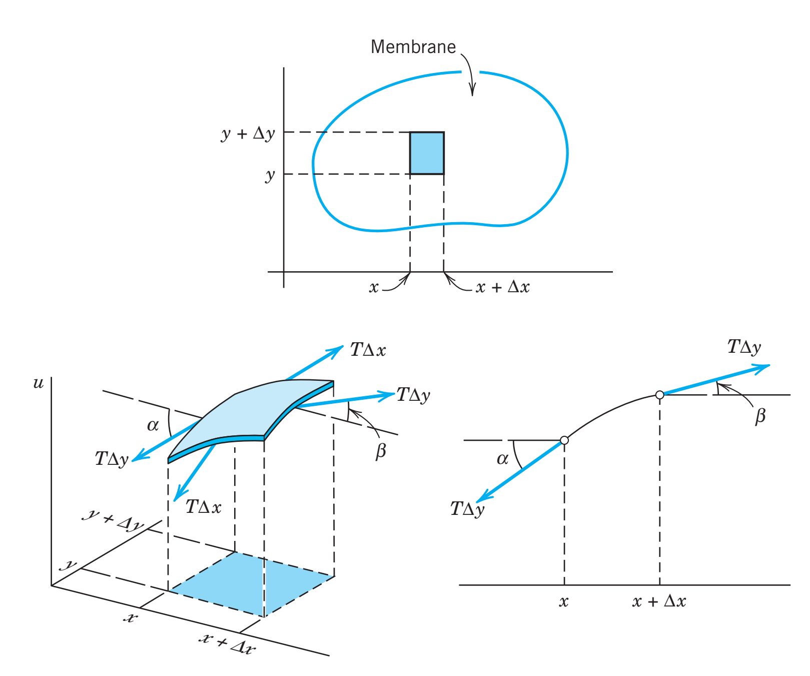


Fig. 301. Vibrating membrane

left side is directed downward. Since the angles are small, we may replace their sines by their tangents. Hence the resultant of those two vertical components is

$$
\begin{aligned}
(1) \qquad T\,\Delta y(\sin\beta - \sin\alpha) &\approx T\,\Delta y(\tan\beta - \tan\alpha) \\
&= T\,\Delta y[u_x(x + \Delta x, y_1) - u_x(x, y_2)]
\end{aligned}
$$

where subscripts x denote partial derivatives and y_1 and y_2 are values between y and $y + \Delta y$. Similarly, the resultant of the vertical components of the forces acting on the other two sides of the portion is

$$
(2) \qquad T\,\Delta x[u_y(x_1, y + \Delta y) - u_y(x_2, y)]
$$

where x_1 and x_2 are values between x and $x + \Delta x$.

Newton's Second Law Gives the PDE of the Model. By Newton's second law (see Sec. 2.4) the sum of the forces given by (1) and (2) is equal to the mass $\rho\,\Delta A$ of that small portion times the acceleration $\partial^2 u/\partial t^2$; here ρ is the mass of the undeflected membrane per unit area, and $\Delta A = \Delta x\,\Delta y$ is the area of that portion when it is undeflected. Thus

$$
\begin{aligned}
\rho\Delta x\,\Delta y\frac{\partial^2 u}{\partial t^2} &= T\,\Delta y[u_x(x + \Delta x, y_1) - u_x(x, y_2)] \\
&\quad + T\,\Delta x[u_y(x_1, y + \Delta y) - u_y(x_2, y)]
\end{aligned}
$$

where the derivative on the left is evaluated at some suitable point $(\tilde{x}, \tilde{y})$ corresponding to that portion. Division by $\rho\Delta x\,\Delta y$ gives

$$\frac{\partial^2 u}{\partial t^2} = \frac{T}{\rho}\left[\frac{u_x(x+\Delta x, y_1) - u_x(x, y_2)}{\Delta x} + \frac{u_y(x_1, y+\Delta y) - u_y(x_2, y)}{\Delta y}\right].$$

If we let Δx and Δy approach zero, we obtain the PDE of the model

(3) $$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \qquad c^2 = \frac{T}{\rho}.$$

This PDE is called the **two-dimensional wave equation**. The expression in parentheses is the Laplacian $\Delta^2 u$ of u (Sec. 10.8). Hence (3) can be written

(3′) $$\frac{\partial^2 u}{\partial t^2} = c^2 \Delta^2 u.$$

Solutions of the wave equation (3) will be obtained and discussed in the next section.

12.9 Rectangular Membrane. Double Fourier Series

Now we develop a solution for the PDE obtained in Sec. 12.8. Details are as follows.

The model of the vibrating membrane for obtaining the displacement $u(x, y, t)$ of a point (x, y) of the membrane from rest $(u = 0)$ at time t is

(1) $$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)$$

(2) $$u = 0 \text{ on the boundary}$$

(3a) $$u(x, y, 0) = f(x, y)$$

(3b) $$u_t(x, y, 0) = g(x, y).$$

Here (1) is the **two-dimensional wave equation** with $c^2 = T/\rho$ just derived, (2) is the **boundary condition** (membrane fixed along the boundary in the xy-plane for all times $t \geqq 0$), and (3) are the **initial conditions** at $t = 0$, consisting of the given *initial displacement* (initial shape) $f(x, y)$ and the given *initial velocity* $g(x, y)$, where $u_t = \partial u/\partial t$. We see that these conditions are quite similar to those for the string in Sec. 12.2.

Fig. 302. Rectangular membrane

Let us consider the **rectangular membrane** R in Fig. 302. This is our first important model. It is much simpler than the circular drumhead, which will follow later. First we note that the boundary in equation (2) is the rectangle in Fig. 302. We shall solve this problem in three steps:

Step 1. By separating variables, first setting $u(x, y, t) = F(x, y)G(t)$ and later $F(x, y) = H(x)Q(y)$ we obtain from (1) an ODE (4) for G and later from a PDE (5) for F two ODEs (6) and (7) for H and Q.

Step 2. From the solutions of those ODEs we determine solutions (13) of (1) ("**eigenfunctions**" u_{mn}) that satisfy the boundary condition (2).

Step 3. We compose the u_{mn} into a double series (14) solving the whole model (1), (2), (3).

Step 1. Three ODEs From the Wave Equation (1)

To obtain ODEs from (1), we apply two successive separations of variables. In the first separation we set $u(x, y, t) = F(x, y)G(t)$. Substitution into (1) gives

$$F\ddot{G} = c^2(F_{xx}G + F_{yy}G)$$

where subscripts denote partial derivatives and dots denote derivatives with respect to t. To separate the variables, we divide both sides by c^2FG:

$$\frac{\ddot{G}}{c^2G} = \frac{1}{F}(F_{xx} + F_{yy}).$$

Since the left side depends only on t, whereas the right side is independent of t, both sides must equal a constant. By a simple investigation we see that only negative values of that constant will lead to solutions that satisfy (2) without being identically zero; this is similar to Sec. 12.3. Denoting that negative constant by $-\nu^2$, we have

$$\frac{\ddot{G}}{c^2G} = \frac{1}{F}(F_{xx} + F_{yy}) = -\nu^2.$$

This gives two equations: for the "**time function**" $G(t)$ we have the ODE

(4) $$\ddot{G} + \lambda^2 G = 0 \qquad \text{where } \lambda = c\nu,$$

and for the "**amplitude function**" $F(x, y)$ a PDE, called the *two-dimensional* **Helmholtz**[3] **equation**

(5) $$F_{xx} + F_{yy} + \nu^2 F = 0.$$

[3]HERMANN VON HELMHOLTZ (1821–1894), German physicist, known for his fundamental work in thermodynamics, fluid flow, and acoustics.

Separation of the Helmholtz equation is achieved if we set $F(x, y) = H(x)Q(y)$. By substitution of this into (5) we obtain

$$\frac{d^2H}{dx^2}Q = -\left(H\frac{d^2Q}{dy^2} + \nu^2 HQ\right).$$

To separate the variables, we divide both sides by HQ, finding

$$\frac{1}{H}\frac{d^2H}{dx^2} = -\frac{1}{Q}\left(\frac{d^2Q}{dy^2} + \nu^2 Q\right).$$

Both sides must equal a constant, by the usual argument. This constant must be negative, say, $-k^2$, because only negative values will lead to solutions that satisfy (2) without being identically zero. Thus

$$\frac{1}{H}\frac{d^2H}{dx^2} = -\frac{1}{Q}\left(\frac{d^2Q}{dy^2} + \nu^2 Q\right) = -k^2.$$

This yields two ODEs for H and Q, namely,

(6) $$\frac{d^2H}{dx^2} + k^2H = 0$$

and

(7) $$\frac{d^2Q}{dy^2} + p^2Q = 0 \qquad \text{where } p^2 = \nu^2 - k^2.$$

Step 2. Satisfying the Boundary Condition

General solutions of (6) and (7) are

$$H(x) = A\cos kx + B\sin kx \qquad \text{and} \qquad Q(y) = C\cos py + D\sin py$$

with constant A, B, C, D. From $u = FG$ and (2) it follows that $F = HQ$ must be zero on the boundary, that is, on the edges $x = 0, x = a, y = 0, y = b$; see Fig. 302. This gives the conditions

$$H(0) = 0, \qquad H(a) = 0, \qquad Q(0) = 0, \qquad Q(b) = 0.$$

Hence $H(0) = A = 0$ and then $H(a) = B\sin ka = 0$. Here we must take $B \neq 0$ since otherwise $H(x) \equiv 0$ and $F(x, y) \equiv 0$. Hence $\sin ka = 0$ or $ka = m\pi$, that is,

$$k = \frac{m\pi}{a} \qquad (m \text{ integer}).$$

In precisely the same fashion we conclude that $C = 0$ and p must be restricted to the values $p = n\pi/b$ where n is an integer. We thus obtain the solutions $H = H_m$, $Q = Q_n$, where

$$H_m(x) = \sin\frac{m\pi x}{a} \quad \text{and} \quad Q_n(y) = \sin\frac{n\pi y}{b}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots. \end{matrix}$$

As in the case of the vibrating string, it is not necessary to consider $m, n = -1, -2, \cdots$ since the corresponding solutions are essentially the same as for positive m and n, expect for a factor -1. Hence the functions

$$F_{mn}(x, y) = H_m(x)Q_n(y) = \sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots, \end{matrix} \tag{8}$$

are solutions of the Helmholtz equation (5) that are zero on the boundary of our membrane.

Eigenfunctions and Eigenvalues. Having taken care of (5), we turn to (4). Since $p^2 = \nu^2 - k^2$ in (7) and $\lambda = c\nu$ in (4), we have

$$\lambda = c\sqrt{k^2 + p^2}.$$

Hence to $k = m\pi/a$ and $p = n\pi/b$ there corresponds the value

$$\lambda = \lambda_{mn} = c\pi\sqrt{\frac{m^2}{a^2} + \frac{n^2}{b^2}}, \qquad \begin{matrix} m = 1, 2, \cdots, \\ n = 1, 2, \cdots, \end{matrix} \tag{9}$$

in the ODE (4). A corresponding general solution of (4) is

$$G_{mn}(t) = B_{mn}\cos\lambda_{mn}t + B^*_{mn}\sin\lambda_{mn}t.$$

It follows that the functions $u_{mn}(x, y, t) = F_{mn}(x, y)G_{mn}(t)$, written out

$$u_{mn}(x, y, t) = (B_{mn}\cos\lambda_{mn}t + B^*_{mn}\sin\lambda_{mn}t)\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b} \tag{10}$$

with λ_{mn} according to (9), are solutions of the wave equation (1) that are zero on the boundary of the rectangular membrane in Fig. 302. These functions are called the **eigenfunctions** or *characteristic functions,* and the numbers λ_{mn} are called the **eigenvalues** or *characteristic values* of the vibrating membrane. The frequency of u_{mn} is $\lambda_{mn}/2\pi$.

Discussion of Eigenfunctions. It is very interesting that, depending on a and b, several functions F_{mn} may correspond to the same eigenvalue. Physically this means that there may exists vibrations having the same frequency but entirely different **nodal lines** (curves of points on the membrane that do not move). Let us illustrate this with the following example.

EXAMPLE 1 Eigenvalues and Eigenfunctions of the Square Membrane

Consider the square membrane with $a = b = 1$. From (9) we obtain its eigenvalues

$$\lambda_{mn} = c\pi\sqrt{m^2 + n^2}. \tag{11}$$

Hence $\lambda_{mn} = \lambda_{nm}$, but for $m \neq n$ the corresponding functions

$$F_{mn} = \sin m\pi x \sin n\pi y \qquad \text{and} \qquad F_{nm} = \sin n\pi x \sin m\pi y$$

are certainly different. For example, to $\lambda_{12} = \lambda_{21} = c\pi\sqrt{5}$ there correspond the two functions

$$F_{12} = \sin \pi x \sin 2\pi y \qquad \text{and} \qquad F_{21} = \sin 2\pi x \sin \pi y.$$

Hence the corresponding solutions

$$u_{12} = (B_{12} \cos c\pi\sqrt{5}t + B^*_{12} \sin c\pi\sqrt{5}t)F_{12} \qquad \text{and} \qquad u_{21} = (B_{21} \cos c\pi\sqrt{5}t + B^*_{21} \sin c\pi\sqrt{5}t)F_{21}$$

have the nodal lines $y = \frac{1}{2}$ and $x = \frac{1}{2}$, respectively (see Fig. 303). Taking $B_{12} = 1$ and $B^*_{12} = B^*_{21} = 0$, we obtain

$$u_{12} + u_{21} = \cos c\pi\sqrt{5}t\,(F_{12} + B_{21}F_{21}) \tag{12}$$

which represents another vibration corresponding to the eigenvalue $c\pi\sqrt{5}$. The nodal line of this function is the solution of the equation

$$F_{12} + B_{21}F_{21} = \sin \pi x \sin 2\pi y + B_{21} \sin 2\pi x \sin \pi y = 0$$

or, since $\sin 2\alpha = 2 \sin \alpha \cos \alpha$,

$$\sin \pi x \sin \pi y(\cos \pi y + B_{21} \cos \pi x) = 0. \tag{13}$$

This solution depends on the value of B_{21} (see Fig. 304).

From (11) we see that even more than two functions may correspond to the same numerical value of λ_{mn}. For example, the four functions F_{18}, F_{81}, F_{47}, and F_{74} correspond to the value

$$\lambda_{18} = \lambda_{81} = \lambda_{47} = \lambda_{74} = c\pi\sqrt{65}, \qquad \text{because} \qquad 1^2 + 8^2 = 4^2 + 7^2 = 65.$$

This happens because 65 can be expressed as the sum of two squares of positive integers in several ways. According to a theorem by Gauss, this is the case for every sum of two squares among whose prime factors there are at least two different ones of the form $4n + 1$ where n is a positive integer. In our case we have $65 = 5 \cdot 13 = (4 + 1)(12 + 1)$. ■

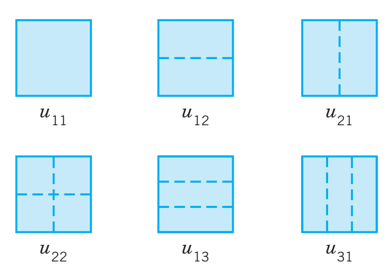


Fig. 303. Nodal lines of the solutions u_{11}, u_{12}, u_{21}, u_{22}, u_{13}, u_{31} in the case of the square membrane

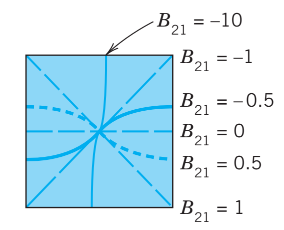


Fig. 304. Nodal lines of the solution (12) for some values of B_{21}

Step 3. Solution of the Model (1), (2), (3). Double Fourier Series

So far we have solutions (10) satisfying (1) and (2) only. To obtain the solutions that also satisfies (3), we proceed as in Sec. 12.3. We consider the double series

$$\textbf{(14)} \qquad \begin{aligned} u(x, y, t) &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} u_{mn}(x, y, t) \\ &= \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} (B_{mn} \cos \lambda_{mn} t + B^*_{mn} \sin \lambda_{mn} t) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} \end{aligned}$$

(without discussing convergence and uniqueness). From (14) and (3a), setting $t = 0$, we have

$$\textbf{(15)} \qquad u(x, y, 0) = \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} B_{mn} \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b} = f(x, y).$$

Suppose that $f(x, y)$ can be represented by (15). (Sufficient for this is the continuity of f, $\partial f/\partial x$, $\partial f/\partial y$, $\partial^2 f/\partial x\, \partial y$ in R.) Then (15) is called the **double Fourier series** of $f(x, y)$. Its coefficients can be determined as follows. Setting

$$(16) \qquad K_m(y) = \sum_{n=1}^{\infty} B_{mn} \sin \frac{n\pi y}{b}$$

we can write (15) in the form

$$f(x, y) = \sum_{m=1}^{\infty} K_m(y) \sin \frac{m\pi x}{a}.$$

For fixed y this is the Fourier sine series of $f(x, y)$, considered as a function of x. From (4) in Sec. 11.3 we see that the coefficients of this expansion are

$$(17) \qquad K_m(y) = \frac{2}{a} \int_0^a f(x, y) \sin \frac{m\pi x}{a}\, dx.$$

Furthermore, (16) is the Fourier sine series of $K_m(y)$, and from (4) in Sec. 11.3 it follows that the coefficients are

$$B_{mn} = \frac{2}{b} \int_0^b K_m(y) \sin \frac{n\pi y}{b}\, dy.$$

From this and (17) we obtain the **generalized Euler formula**

$$\textbf{(18)} \qquad B_{mn} = \frac{4}{ab} \int_0^b \int_0^a f(x, y) \sin \frac{m\pi x}{a} \sin \frac{n\pi y}{b}\, dx\, dy \qquad \begin{aligned} m &= 1, 2, \cdots \\ n &= 1, 2, \cdots \end{aligned}$$

for the **Fourier coefficients** of $f(x, y)$ in the double Fourier series (15).

The B_{mn} in (14) are now determined in terms of $f(x, y)$. To determine the B^*_{mn}, we differentiate (14) termwise with respect to t; using (3b), we obtain

$$\left.\frac{\partial u}{\partial t}\right|_{t=0} = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} B^*_{mn}\lambda_{mn}\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{b} = g(x, y).$$

Suppose that $g(x, y)$ can be developed in this double Fourier series. Then, proceeding as before, we find that the coefficients are

$$\textbf{(19)} \qquad B^*_{mn} = \frac{4}{ab\lambda_{mn}}\int_0^b\int_0^a g(x, y)\sin\frac{m\pi x}{a}\sin\frac{n\pi y}{a}\,dx\,dy \qquad \begin{matrix} m = 1, 2, \cdots \\ n = 1, 2, \cdots . \end{matrix}$$

Result. *If f and g in* (3) *are such that u can be represented by* (14), *then* (14) *with coefficients* (18) *and* (19) *is the solution of the model* (1), (2), (3).

EXAMPLE 2 Vibration of a Rectangular Membrane

Find the vibrations of a rectangular membrane of sides $a = 4$ ft and $b = 2$ ft (Fig. 305) if the tension is 12.5 lb/ft, the density is 2.5 slugs/ft^2 (as for light rubber), the initial velocity is 0, and the initial displacement is

$$\text{(20)} \qquad f(x, y) = 0.1(4x - x^2)(2y - y^2) \text{ ft.}$$

Fig. 305. Example 2

Solution. $c^2 = T/\rho = 12.5/2.5 = 5$ [ft^2/sec^2]. Also $B^*_{mn} = 0$ from (19). From (18) and (20),

$$\begin{aligned} B_{mn} &= \frac{4}{4\cdot 2}\int_0^2\int_0^4 0.1(4x - x^2)(2y - y^2)\sin\frac{m\pi x}{4}\sin\frac{n\pi y}{2}\,dx\,dy \\ &= \frac{1}{20}\int_0^4 (4x - x^2)\sin\frac{m\pi x}{4}\,dx\int_0^2 (2y - y^2)\sin\frac{n\pi y}{2}\,dy. \end{aligned}$$

Two integrations by parts give for the first integral on the right

$$\frac{128}{m^3\pi^3}[1 - (-1)^m] = \frac{256}{m^3\pi^3} \qquad (m \text{ odd})$$

and for the second integral

$$\frac{16}{n^3\pi^3}[1 - (-1)^n] = \frac{32}{n^3\pi^3} \qquad (n \text{ odd}).$$

For even m or n we get 0. Together with the factor 1/20 we thus have $B_{mn} = 0$ if m or n is even and

$$B_{mn} = \frac{256 \cdot 32}{20m^3n^3\pi^6} \approx \frac{0.426050}{m^3n^3} \qquad (m \text{ and } n \text{ both odd}).$$

From this, (9), and (14) we obtain the answer

$$u(x, y, t) = 0.426050 \sum_{m,n \text{ odd}} \sum \frac{1}{m^3n^3} \cos\left(\frac{\sqrt{5}\pi}{4}\sqrt{m^2 + 4n^2}\right) t \sin\frac{m\pi x}{4} \sin\frac{n\pi y}{2}$$

$$(21) \qquad = 0.426050\left(\cos\frac{\sqrt{5}\pi\sqrt{5}}{4} t \sin\frac{\pi x}{4} \sin\frac{\pi y}{2} + \frac{1}{27}\cos\frac{\sqrt{5}\pi\sqrt{37}}{4} t \sin\frac{\pi x}{4} \sin\frac{3\pi y}{2}\right.$$
$$\left. + \frac{1}{27}\cos\frac{\sqrt{5}\pi\sqrt{13}}{4} t \sin\frac{3\pi x}{4} \sin\frac{\pi y}{2} + \frac{1}{729}\cos\frac{\sqrt{5}\pi\sqrt{45}}{4} t \sin\frac{3\pi x}{4} \sin\frac{3\pi y}{2} + \cdots\right).$$

To discuss this solution, we note that the first term is very similar to the initial shape of the membrane, has no nodal lines, and is by far the dominating term because the coefficients of the next terms are much smaller. The second term has two horizontal nodal lines ($y = \frac{2}{3}, \frac{4}{3}$), the third term two vertical ones ($x = \frac{4}{3}, \frac{8}{3}$), the fourth term two horizontal and two vertical ones, and so on. ■

PROBLEM SET 12.9

1. **Frequency.** How does the frequency of the eigenfunctions of the rectangular membrane change (a) If we double the tension? (b) If we take a membrane of half the density of the original one? (c) If we double the sides of the membrane? Give reasons.
2. **Assumptions.** Which part of Assumption 2 cannot be satisfied exactly? Why did we also assume that the angles of inclination are small?
3. Determine and sketch the nodal lines of the square membrane for $m = 1, 2, 3, 4$ and $n = 1, 2, 3, 4$.

4–8 DOUBLE FOURIER SERIES

Represent $f(x, y)$ by a series (15), where

4. $f(x, y) = 1, \quad a = b = 1$
5. $f(x, y) = y, \quad a = b = 1$
6. $f(x, y) = x, \quad a = b = 1$
7. $f(x, y) = xy, \quad a$ and b arbitrary
8. $f(x, y) = xy(a - x)(b - y), \quad a$ and b arbitrary
9. **CAS PROJECT. Double Fourier Series. (a)** Write a program that gives and graphs partial sums of (15). Apply it to Probs. 5 and 6. Do the graphs show that those partial sums satisfy the boundary condition (3a)? Explain why. Why is the convergence rapid?

   (b) Do the tasks in (a) for Prob. 4. Graph a portion, say, $0 < x < \frac{1}{2}, 0 < y < \frac{1}{2}$, of several partial sums on common axes, so that you can see how they differ. (See Fig. 306.)

   (c) Do the tasks in (b) for functions of your choice.

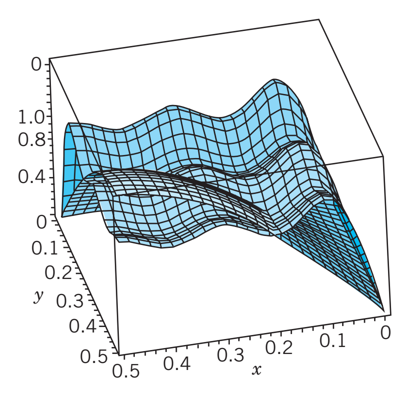


Fig. 306. Partial sums $S_{2,2}$ and $S_{10,10}$ in CAS Project 9b

10. **CAS EXPERIMENT. Quadruples of F_{mn}.** Write a program that gives you four numerically equal λ_{mn} in Example 1, so that four different F_{mn} correspond to it. Sketch the nodal lines of F_{18}, F_{81}, F_{47}, F_{74} in Example 1 and similarly for further F_{mn} that you will find.

11–13 SQUARE MEMBRANE

Find the deflection $u(x, y, t)$ of the square membrane of side π and $c^2 = 1$ for initial velocity 0 and initial deflection

11. $0.1 \sin 2x \sin 4y$
12. $0.01 \sin x \sin y$
13. $0.1\,xy(\pi - x)(\pi - y)$

14–19 RECTANGULAR MEMBRANE

14. Verify the discussion of (21) in Example 2.

15. Do Prob. 3 for the membrane with $a = 4$ and $b = 2$.

16. Verify B_{mn} in Example 2 by integration by parts.

17. Find eigenvalues of the rectangular membrane of sides $a = 2$ and $b = 1$ to which there correspond two or more different (independent) eigenfunctions.

18. Minimum property. Show that among all rectangular membranes of the same area $A = ab$ and the same c the square membrane is that for which u_{11} [see (10)] has the lowest frequency.

19. Deflection. Find the deflection of the membrane of sides a and b with $c^2 = 1$ for the initial deflection

$$f(x, y) = \sin \frac{6\pi x}{a} \sin \frac{2\pi y}{b} \text{ and initial velocity 0.}$$

20. Forced vibrations. Show that forced vibrations of a membrane are modeled by the PDE $u_{tt} = c^2\nabla^2 u + P/\rho$, where $P(x, y, t)$ is the external force per unit area acting perpendicular to the xy-plane.

12.10 Laplacian in Polar Coordinates. Circular Membrane. Fourier–Bessel Series

It is a ***general principle*** in boundary value problems for PDEs to *choose coordinates that make the formula for the boundary as simple as possible*. Here polar coordinates are used for this purpose as follows. Since we want to discuss circular membranes (drumheads), we first transform the Laplacian in the wave equation (1), Sec. 12.9,

$$u_{tt} = c^2\nabla^2 u = c^2(u_{xx} + u_{yy}) \tag{1}$$

(subscripts denoting partial derivatives) into **polar coordinates** r, θ defined by $x = r\cos\theta$, $y = r\sin\theta$; thus,

$$r = \sqrt{x^2 + y^2}, \qquad \tan\theta = \frac{y}{x}.$$

By the chain rule (Sec. 9.6) we obtain

$$u_x = u_r r_x + u_\theta \theta_x.$$

Differentiating once more with respect to x and using the product rule and then again the chain rule gives

$$\begin{aligned} u_{xx} &= (u_r r_x)_x + (u_\theta \theta_x)_x \\ &= (u_r)_x r_x + u_r r_{xx} + (u_\theta)_x \theta_x + u_\theta \theta_{xx} \\ &= (u_{rr} r_x + u_{r\theta}\theta_x) r_x + u_r r_{xx} + (u_{\theta r} r_x + u_{\theta\theta}\theta_x)\theta_x + u_\theta \theta_{xx}. \end{aligned} \tag{2}$$

Also, by differentiation of r and θ we find

$$r_x = \frac{x}{\sqrt{x^2 + y^2}} = \frac{x}{r}, \qquad \theta_x = \frac{1}{1 + (y/x)^2}\left(-\frac{y}{x^2}\right) = -\frac{y}{r^2}.$$

Differentiating these two formulas again, we obtain

$$r_{xx} = \frac{r - xr_x}{r^2} = \frac{1}{r} - \frac{x^2}{r^3} = \frac{y^2}{r^3}, \qquad \theta_{xx} = -y\left(-\frac{2}{r^3}\right)r_x = \frac{2xy}{r^4}.$$

We substitute all these expressions into (2). Assuming continuity of the first and second partial derivatives, we have $u_{r\theta} = u_{\theta r}$, and by simplifying,

$$u_{xx} = \frac{x^2}{r^2}u_{rr} - 2\frac{xy}{r^3}u_{r\theta} + \frac{y^2}{r^4}u_{\theta\theta} + \frac{y^2}{r^3}u_r + 2\frac{xy}{r^4}u_\theta. \tag{3}$$

In a similar fashion it follows that

$$u_{yy} = \frac{y^2}{r^2}u_{rr} + 2\frac{xy}{r^3}u_{r\theta} + \frac{x^2}{r^4}u_{\theta\theta} + \frac{x^2}{r^3}u_r - 2\frac{xy}{r^4}u_\theta. \tag{4}$$

By adding (3) and (4) we see that the **Laplacian of *u* in polar coordinates** is

$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}. \tag{5}$$

Circular Membrane

Circular membranes are important parts of drums, pumps, microphones, telephones, and other devices. This accounts for their great importance in engineering. Whenever a circular membrane is plane and its material is elastic, but offers no resistance to bending (this excludes thin metallic membranes!), its vibrations are modeled by the **two-dimensional wave equation in polar coordinates** obtained from (1) with $\nabla^2 u$ given by (5), that is,

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) \qquad c^2 = \frac{T}{\rho}. \tag{6}$$

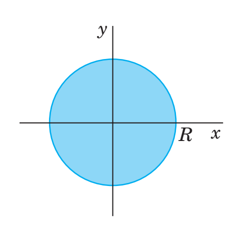


Fig. 307. Circular membrane

We shall consider a membrane of radius R (Fig. 307) and determine solutions $u(r, t)$ that are radially symmetric. (Solutions also depending on the angle θ will be discussed in the problem set.) Then $u_{\theta\theta} = 0$ in (6) and the model of the problem (the analog of (1), (2), (3) in Sec. 12.9) is

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r}\right) \tag{7}$$

$$u(R, t) = 0 \text{ for all } t \geqq 0 \tag{8}$$

$$u(r, 0) = f(r) \tag{9a}$$

$$u_t(r, 0) = g(r). \tag{9b}$$

Here (8) means that the membrane is fixed along the boundary circle $r = R$. The initial deflection $f(r)$ and the initial velocity $g(r)$ depend only on r, not on θ, so that we can expect radially symmetric solutions $u(r, t)$.

Step 1. Two ODEs From the Wave Equation (7). Bessel's Equation

Using the **method of separation of variables**, we first determine solutions $u(r, t) = W(r)G(t)$. (We write W, not F because W depends on r, whereas F, used before, depended on x.) Substituting $u = WG$ and its derivatives into (7) and dividing the result by c^2WG, we get

$$\frac{\ddot{G}}{c^2G} = \frac{1}{W}\left(W'' + \frac{1}{r}W'\right)$$

where dots denote derivatives with respect to t and primes denote derivatives with respect to r. The expressions on both sides must equal a constant. This constant must be negative, say, $-k^2$, in order to obtain solutions that satisfy the boundary condition without being identically zero. Thus,

$$\frac{\ddot{G}}{c^2G} = \frac{1}{W}\left(W'' + \frac{1}{r}W'\right) = -k^2.$$

This gives the two linear ODEs

(10) $$\ddot{G} + \lambda^2 G = 0 \qquad \text{where } \lambda = ck$$

and

(11) $$W'' + \frac{1}{r}W' + k^2W = 0.$$

We can reduce (11) to Bessel's equation (Sec. 5.4) if we set $s = kr$. Then $1/r = k/s$ and, retaining the notation W for simplicity, we obtain by the chain rule

$$W' = \frac{dW}{dr} = \frac{dW}{ds}\frac{ds}{dr} = \frac{dW}{ds}k \qquad \text{and} \qquad W'' = \frac{d^2W}{ds^2}k^2.$$

By substituting this into (11) and omitting the common factor k^2 we have

(12) $$\frac{d^2W}{ds^2} + \frac{1}{s}\frac{dW}{ds} + W = 0.$$

This is **Bessel's equation** (1), Sec. 5.4, with parameter $\nu = 0$.

Step 2. Satisfying the Boundary Condition (8)

Solutions of (12) are the Bessel functions J_0 and Y_0 of the first and second kind (see Secs. 5.4, 5.5). But Y_0 becomes infinite at 0, so that we cannot use it because the deflection of the membrane must always remain finite. This leaves us with

(13) $$W(r) = J_0(s) = J_0(kr) \qquad (s = kr).$$

On the boundary $r = R$ we get $W(R) = J_0(kR) = 0$ from (8) (because $G \equiv 0$ would imply $u \equiv 0$). We can satisfy this condition because J_0 has (infinitely many) positive zeros, $s = \alpha_1, \alpha_2, \cdots$ (see Fig. 308), with numerical values

$$\alpha_1 = 2.4048, \quad \alpha_2 = 5.5201, \quad \alpha_3 = 8.6537, \quad \alpha_4 = 11.7915, \quad \alpha_5 = 14.9309$$

and so on. (For further values, consult your CAS or Ref. [GenRef1] in App. 1.) These zeros are slightly irregularly spaced, as we see. Equation (13) now implies

$$(14) \qquad kR = \alpha_m \quad \text{thus} \quad k = k_m = \frac{\alpha_m}{R}, \qquad m = 1, 2, \cdots .$$

Hence the functions

$$(15) \qquad W_m(r) = J_0(k_m r) = J_0\left(\frac{\alpha_m}{R} r\right), \qquad m = 1, 2, \cdots$$

are solutions of (11) that are zero on the boundary circle $r = R$.

Eigenfunctions and Eigenvalues. For W_m in (15), a corresponding general solution of (10) with $\lambda = \lambda_m = ck_m = c\alpha_m/R$ is

$$G_m(t) = A_m \cos \lambda_m t + B_m \sin \lambda_m t.$$

Hence the functions

$$\textbf{(16)} \qquad u_m(r, t) = W_m(r)G_m(t) = (A_m \cos \lambda_m t + B_m \sin \lambda_m t)J_0(k_m r)$$

with $m = 1, 2, \cdots$ are solutions of the wave equation (7) satisfying the boundary condition (8). These are the **eigenfunctions** of our problem. The corresponding **eigenvalues** are λ_m.

The vibration of the membrane corresponding to u_m is called the mth **normal mode**; it has the frequency $\lambda_m/2\pi$ cycles per unit time. Since the zeros of the Bessel function J_0 are not regularly spaced on the axis (in contrast to the zeros of the sine functions appearing in the case of the vibrating string), the sound of a drum is entirely different from that of a violin. The forms of the normal modes can easily be obtained from Fig. 308 and are shown in Fig. 309. For $m = 1$, all the points of the membrane move up (or down) at the same time. For $m = 2$, the situation is as follows. The function $W_2(r) = J_0(\alpha_2 r/R)$ is zero for $\alpha_2 r/R = \alpha_1$, thus $r = \alpha_1 R/\alpha_2$. The circle $r = \alpha_1 R/\alpha_2$ is, therefore, **nodal line**, and when at some instant the central part of the membrane moves up, the outer part ($r > \alpha_1 R/\alpha_2$) moves down, and conversely. The solution $u_m(r, t)$ has $m - 1$ nodal lines, which are circles (Fig. 309).

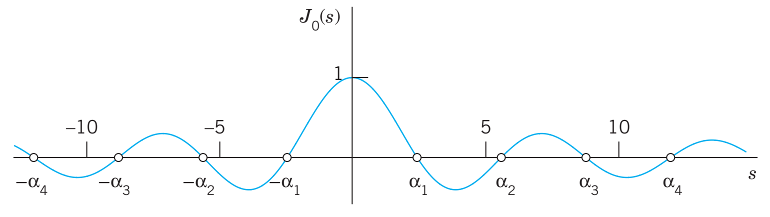


Fig. 308. Bessel function $J_0(s)$

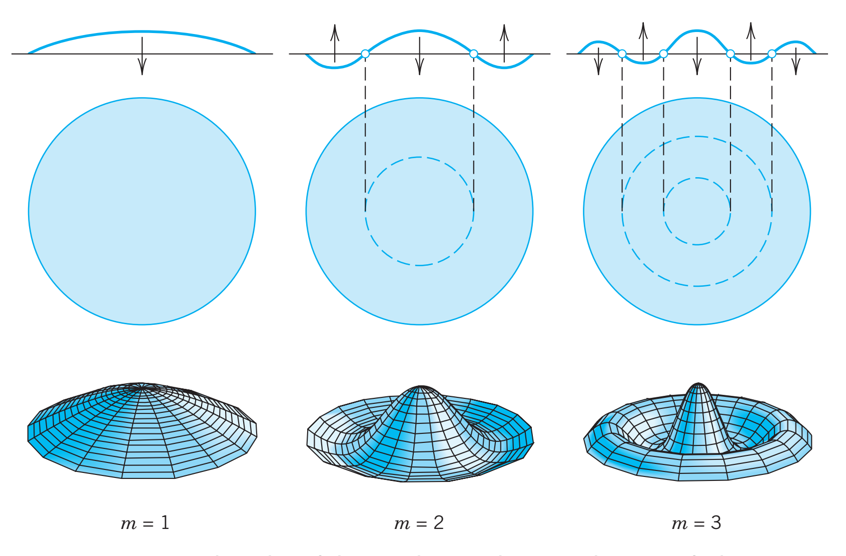


Fig. 309. Normal modes of the circular membrane in the case of vibrations independent of the angle

Step 3. Solution of the Entire Problem

To obtain a solution $u(r, t)$ that also satisfies the initial conditions (9), we may proceed as in the case of the string. That is, we consider the series

$$\textbf{(17)} \qquad u(r, t) = \sum_{m=1}^{\infty} W_m(r)G_m(t) = \sum_{m=1}^{\infty} (A_m \cos \lambda_m t + B_m \sin \lambda_m t) J_0\left(\frac{\alpha_m}{R} r\right)$$

(leaving aside the problems of convergence and uniqueness). Setting $t = 0$ and using (9a), we obtain

$$(18) \qquad u(r, 0) = \sum_{m=1}^{\infty} A_m J_0\left(\frac{\alpha_m}{R} r\right) = f(r).$$

Thus for the series (17) to satisfy the condition (9a), the constants A_m must be the coefficients of the **Fourier–Bessel series** (18) that represents $f(r)$ in terms of $J_0(\alpha_m r/R)$; that is [see (9) in Sec. 11.6 with $n = 0$, $\alpha_{0,m} = \alpha_m$, and $x = r$],

$$\textbf{(19)} \qquad A_m = \frac{2}{R^2 J_1^2(\alpha_m)} \int_0^R r f(r) J_0\left(\frac{\alpha_m}{R} r\right) dr \qquad (m = 1, 2, \cdots).$$

Differentiability of $f(r)$ in the interval $0 \leqq r \leqq R$ is sufficient for the existence of the development (18); see Ref. [A13]. The coefficients B_m in (17) can be determined from (9b) in a similar fashion. Numeric values of A_m and B_m may be obtained from a CAS or by a numeric integration method, using tables of J_0 and J_1. However, numeric integration can sometimes be ***avoided***, as the following example shows.

EXAMPLE 1 Vibrations of a Circular Membrane

Find the vibrations of a circular drumhead of radius 1 ft and density 2 slugs/ft^2 if the tension is 8 lb/ft, the initial velocity is 0, and the initial displacement is

$$f(r) = 1 - r^2 \text{ [ft]}.$$

Solution. $c^2 = T/\rho = \frac{8}{2} = 4$ [ft^2/sec^2]. Also $B_m = 0$, since the initial velocity is 0. From (10) in Sec. 11.6, since $R = 1$, we obtain

$$\begin{aligned} A_m &= \frac{2}{J_1^2(\alpha_m)} \int_0^1 r(1 - r^2) J_0(\alpha_m r)\, dr \\ &= \frac{4J_2(\alpha_m)}{\alpha_m^2 J_1^2(\alpha_m)} \\ &= \frac{8}{\alpha_m^3 J_1(\alpha_m)} \end{aligned}$$

where the last equality follows from (21c), Sec. 5.4, with $\nu = 1$, that is,

$$J_2(\alpha_m) = \frac{2}{\alpha_m} J_1(\alpha_m) - J_0(\alpha_m) = \frac{2}{\alpha_m} J_1(\alpha_m).$$

Table 9.5 on p. 409 of [GenRef1] gives α_m and $J_0'(\alpha_m)$. From this we get $J_1(\alpha_m) = -J_0'(\alpha_m)$ by (21b), Sec. 5.4, with $\nu = 0$, and compute the coefficients A_m:

m	α_m	$J_1(\alpha_m)$	$J_2(\alpha_m)$	A_m
1	2.40483	0.51915	0.43176	1.10801
2	5.52008	−0.34026	−0.12328	−0.13978
3	8.65373	0.27145	0.06274	0.04548
4	11.79153	−0.23246	−0.03943	−0.02099
5	14.93092	0.20655	0.02767	0.01164
6	18.07106	−0.18773	−0.02078	−0.00722
7	21.21164	0.17327	0.01634	0.00484
8	24.35247	−0.16170	−0.01328	−0.00343
9	27.49348	0.15218	0.01107	0.00253
10	30.63461	−0.14417	−0.00941	−0.00193

Thus

$$f(r) = 1.108 J_0(2.4048r) - 0.140 J_0(5.5201r) + 0.045 J_0(8.6537r) - \cdots.$$

We see that the coefficients decrease relatively slowly. The sum of the explicitly given coefficients in the table is 0.99915. The sum of *all* the coefficients should be 1. (Why?) Hence by the Leibniz test in App. A3.3 the partial sum of those terms gives about three correct decimals of the amplitude $f(r)$.

Since

$$\lambda_m = ck_m = c\alpha_m/R = 2\alpha_m,$$

from (17) we thus obtain the solution (with r measured in feet and t in seconds)

$$u(r, t) = 1.108 J_0(2.4048r) \cos 4.8097t - 0.140 J_0(5.5201r) \cos 11.0402t + 0.045 J_0(8.6537r) \cos 17.3075t - \cdots.$$

In Fig. 309, $m = 1$ gives an idea of the motion of the first term of our series, $m = 2$ of the second term, and $m = 3$ of the third term, so that we can "see" our result about as well as for a violin string in Sec. 12.3. ■

PROBLEM SET 12.10

1–3 RADIAL SYMMETRY

1. Why did we introduce polar coordinates in this section?

2. Radial symmetry reduces (5) to $\nabla^2 u = u_{rr} + u_r/r$. Derive this directly from $\nabla^2 u = u_{xx} + u_{yy}$. Show that the only solution of $\nabla^2 u = 0$ depending only on $r = \sqrt{x^2 + y^2}$ is $u = a \ln r + b$ with arbitrary constants a and b.

3. Alternative form of (5). Show that (5) can be written $\nabla^2 u = (ru_r)_r/r + u_{\theta\theta}/r^2$, a form that is often practical.

BOUNDARY VALUE PROBLEMS. SERIES

4. TEAM PROJECT. Series for Dirichlet and Neumann Problems

(a) Show that $u_n = r^n \cos n\theta$, $u_n = r^n \sin n\theta$, $n = 0, 1, \cdots$, are solutions of Laplace's equation $\nabla^2 u = 0$ with $\nabla^2 u$ given by (5). (What would u_n be in Cartesian coordinates? Experiment with small n.)

(b) Dirichlet problem (See Sec. 12.6) Assuming that termwise differentiation is permissible, show that a solution of the Laplace equation in the disk $r < R$ satisfying the boundary condition $u(R, \theta) = f(\theta)$ (R and f given) is

$$u(r, \theta) = a_0 + \sum_{n=1}^{\infty} \left[a_n \left(\frac{r}{R}\right)^n \cos n\theta + b_n \left(\frac{r}{R}\right)^n \sin n\theta \right] \quad (20)$$

where a_n, b_n are the Fourier coefficients of f (see Sec. 11.1).

(c) Dirichlet problem. Solve the Dirichlet problem using (20) if $R = 1$ and the boundary values are $u(\theta) = -100$ volts if $-\pi < \theta < 0$, $u(\theta) = 100$ volts if $0 < \theta < \pi$. (Sketch this disk, indicate the boundary values.)

(d) Neumann problem. Show that the solution of the Neumann problem $\nabla^2 u = 0$ if $r < R$, $u_N(R, \theta) = f(\theta)$ (where $u_N = \partial u/\partial N$ is the directional derivative in the direction of the outer normal) is

$$u(r, \theta) = A_0 + \sum_{n=1}^{\infty} r^n (A_n \cos n\theta + B_n \sin n\theta)$$

with arbitrary A_0 and

$$A_n = \frac{1}{\pi n R^{n-1}} \int_{-\pi}^{\pi} f(\theta) \cos n\theta \, d\theta,$$

$$B_n = \frac{1}{\pi n R^{n-1}} \int_{-\pi}^{\pi} f(\theta) \sin n\theta \, d\theta.$$

(e) Compatibility condition. Show that (9), Sec. 10.4, imposes on $f(\theta)$ in (d) the *"compatibility condition"*

$$\int_{-\pi}^{\pi} f(\theta) \, d\theta = 0.$$

(f) Neumann problem. Solve $\nabla^2 u = 0$ in the annulus $1 < r < 2$ if $u_r(1, \theta) = \sin \theta$, $u_r(2, \theta) = 0$.

5–8 ELECTROSTATIC POTENTIAL. STEADY-STATE HEAT PROBLEMS

The electrostatic potential satisfies Laplace's equation $\nabla^2 u = 0$ in any region free of charges. Also the heat equation $u_t = c^2 \nabla^2 u$ (Sec. 12.5) reduces to Laplace's equation if the temperature u is time-independent ("**steady-state case**"). Using (20), find the potential (equivalently: the steady-state temperature) in the disk $r < 1$ if the boundary values are (sketch them, to see what is going on).

5. $u(1, \theta) = 220$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise

6. $u(1, \theta) = 400 \cos^3 \theta$

7. $u(1, \theta) = 110|\theta|$ if $-\pi < \theta < \pi$

8. $u(1, \theta) = \theta$ if $-\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise

9. CAS EXPERIMENT. Equipotential Lines. Guess what the equipotential lines $u(r, \theta) = \text{const}$ in Probs. 5 and 7 may look like. Then graph some of them, using partial sums of the series.

10. Semidisk. Find the electrostatic potential in the semidisk $r < 1$, $0 < \theta < \pi$ which equals $110\theta(\pi - \theta)$ on the semicircle $r = 1$ and 0 on the segment $-1 < x < 1$.

11. Semidisk. Find the steady-state temperature in a semicircular thin plate $r < a$, $0 < \theta < \pi$ with the semicircle $r = a$ kept at constant temperature u_0 and the segment $-a < x < a$ at 0.

CIRCULAR MEMBRANE

12. CAS PROJECT. Normal Modes. (a) Graph the normal modes u_4, u_5, u_6 as in Fig. 306.

(b) Write a program for calculating the A_m's in Example 1 and extend the table to $m = 15$. Verify numerically that $\alpha_m \approx (m - \frac{1}{4})\pi$ and compute the error for $m = 1, \cdots, 10$.

(c) Graph the initial deflection $f(r)$ in Example 1 as well as the first three partial sums of the series. Comment on accuracy.

(d) Compute the radii of the nodal lines of u_2, u_3, u_4 when $R = 1$. How do these values compare to those of the nodes of the vibrating string of length 1? Can you establish any empirical laws by experimentation with further u_m?

13. Frequency. What happens to the frequency of an eigenfunction of a drum if you double the tension?

14. Size of a drum. A small drum should have a higher fundamental frequency than a large one, tension and density being the same. How does this follow from our formulas?

15. Tension. Find a formula for the tension required to produce a desired fundamental frequency f_1 of a drum.

16. Why is $A_1 + A_2 + \cdots = 1$ in Example 1? Compute the first few partial sums until you get 3-digit accuracy. What does this problem mean in the field of music?

17. Nodal lines. Is it possible that for fixed c and R two or more u_m [see (16)] with different nodal lines correspond to the same eigenvalue? (Give a reason.)

18. Nonzero initial velocity is more of theoretical interest because it is difficult to obtain experimentally. Show that for (17) to satisfy (9b) we must have

$$B_m = K_m \int_0^R rg(r)J_0(\alpha_m r/R)\, dr \tag{21}$$

where $K_m = 2/(c\alpha_m R)J_1^2(\alpha_m)$.

VIBRATIONS OF A CIRCULAR MEMBRANE DEPENDING ON BOTH r AND θ

19. (Separations) Show that substitution of $u = F(r, \theta)G(t)$ into the wave equation (6), that is,

$$u_{tt} = c^2\left(u_{rr} + \frac{1}{r}u_r + \frac{1}{r_2}u_{\theta\theta}\right), \tag{22}$$

gives an ODE and a PDE

$$\ddot{G} + \lambda^2 G = 0, \qquad \text{where } \lambda = ck, \tag{23}$$

$$F_{rr} + \frac{1}{r}F_r + \frac{1}{r^2}F_{\theta\theta} + k^2F = 0. \tag{24}$$

Show that the PDE can now be separated by substituting $F = W(r)Q(\theta)$, giving

$$Q'' + n^2Q = 0, \tag{25}$$

$$r^2W'' + rW' + (k^2r^2 - n^2)W = 0. \tag{26}$$

20 Periodicity. Show that $Q(\theta)$ must be periodic with period 2π and, therefore, $n = 0, 1, 2, \cdots$ in (25) and (26). Show that this yields the solutions $Q_n = \cos n\theta$, $Q_n^* = \sin n\theta$, $W_n = J_n(kr)$, $n = 0, 1, \cdots$.

21. Boundary condition. Show that the boundary condition

$$u(R, \theta, t) = 0 \tag{27}$$

leads to $k = k_{mn} = \alpha_{mn}/R$, where $s = \alpha_{nm}$ is the mth positive zero of $J_n(s)$.

22. Solutions depending on both r and θ. Show that solutions of (22) satisfying (27) are (see Fig. 310)

$$\begin{aligned} u_{nm} &= (A_{nm}\cos ck_{nm}t + B_{nm}\sin ck_{nm}t) \\ &\quad \times J_n(k_{nm}r)\cos n\theta \\ u_{nm}^* &= (A_{nm}^*\cos ck_{nm}t + B_{nm}^*\sin ck_{nm}t) \\ &\quad \times J_n(k_{nm}r)\sin n\theta \end{aligned} \tag{28}$$

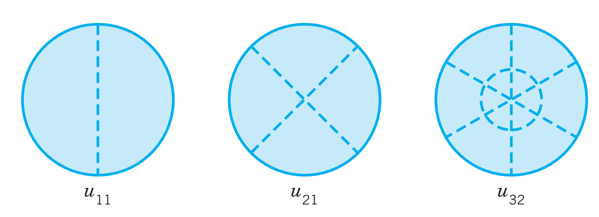


Fig. 310. Nodal lines of some of the solutions (28)

23. Initial condition. Show that $u_t(r, \theta, 0) = 0$ gives $B_{nm} = 0$, $B_{nm}^* = 0$ in (28).

24. Show that $u_{0m}^* = 0$ and u_{0m} is identical with (16) in this section.

25. Semicircular membrane. Show that u_{11} represents the fundamental mode of a semicircular membrane and find the corresponding frequency when $c^2 = 1$ and $R = 1$.

12.11 Laplace's Equation in Cylindrical and Spherical Coordinates. Potential

One of the most important PDEs in physics and engineering applications is **Laplace's equation**, given by

(1) $$\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0.$$

Here, x, y, z are Cartesian coordinates in space (Fig. 167 in Sec. 9.1), $u_{xx} = \partial^2 u/\partial x^2$, etc. The expression $\nabla^2 u$ is called the **Laplacian** of u. The theory of the solutions of (1) is called **potential theory**. Solutions of (1) that have ***continuous*** second partial derivatives are known as **harmonic functions**.

Laplace's equation occurs mainly in **gravitation**, **electrostatics** (see Theorem 3, Sec. 9.7), steady-state **heat flow** (Sec. 12.5), and **fluid flow** (to be discussed in Sec. 18.4).

Recall from Sec. 9.7 that the gravitational **potential** $u(x, y, z)$ at a point (x, y, z) resulting from a single mass located at a point (X, Y, Z) is

(2) $$u(x, y, z) = \frac{c}{r} = \frac{c}{\sqrt{(x - X)^2 + (y - Y)^2 + (z - Z)^2}} \qquad (r > 0)$$

and u satisfies (1). Similarly, if mass is distributed in a region T in space with density $\rho(X, Y, Z)$, its potential at a point (x, y, z) not occupied by mass is

(3) $$u(x, y, z) = k \iiint_T \frac{\rho(X, Y, Z)}{r}\, dX\, dY\, dZ.$$

It satisfies (1) because $\nabla^2(1/r) = 0$ (Sec. 9.7) and ρ is not a function of x, y, z.

Practical problems involving Laplace's equation are boundary value problems in a region T in space with boundary surface S. Such problems can be grouped into three types (see also Sec. 12.6 for the two-dimensional case):

(I) **First boundary value problem or Dirichlet problem** if u is prescribed on S.

(II) **Second boundary value problem** or **Neumann problem** if the normal derivative $u_n = \partial u/\partial n$ is prescribed on S.

(III) **Third** or **mixed boundary value problem** or **Robin problem** if u is prescribed on a portion of S and u_n on the remaining portion of S.

In general, when we want to solve a boundary value problem, we have to first select the appropriate coordinates in which the boundary surface S has a simple representation. Here are some examples followed by some applications.

Laplacian in Cylindrical Coordinates

The first step in solving a boundary value problem is generally the introduction of coordinates in which the boundary surface S has a simple representation. Cylindrical symmetry (a cylinder as a region T) calls for cylindrical coordinates r, θ, z related to x, y, z by

(4) $$x = r\cos\theta, \qquad y = r\sin\theta, \qquad z = z \qquad \text{(Fig. 311).}$$

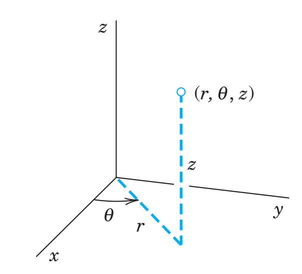


Fig. 311. Cylindrical coordinates
($r \geqq 0, 0 \leqq \theta \leqq 2\pi$)

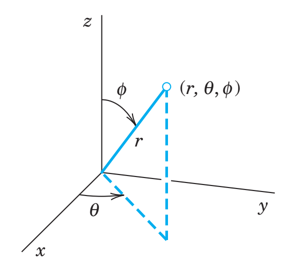


Fig. 312. Spherical coordinates
($r \geqq 0, 0 \leqq \theta \leqq 2\pi, 0 \leqq \varphi \leqq \pi$)

For these we get $\nabla^2 u$ immediately by adding u_{zz} to (5) in Sec. 12.10; thus,

(5)
$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2} + \frac{\partial^2 u}{\partial z^2}.$$

Laplacian in Spherical Coordinates

Spherical symmetry (a ball as region T bounded by a sphere S) requires **spherical coordinates** r, θ, ϕ related to x, y, z by

(6) $$x = r\cos\theta\,\sin\phi, \qquad y = r\sin\theta\,\sin\phi, \qquad z = r\cos\phi \qquad \text{(Fig. 312).}$$

Using the chain rule (as in Sec. 12.10), we obtain $\nabla^2 u$ in spherical coordinates

(7)
$$\nabla^2 u = \frac{\partial^2 u}{\partial r^2} + \frac{2}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \phi^2} + \frac{\cot\phi}{r^2}\frac{\partial u}{\partial \phi} + \frac{1}{r^2\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}.$$

We leave the details as an exercise. It is sometimes practical to write (7) in the form

(7′) $$\nabla^2 u = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\,\frac{\partial u}{\partial \phi}\right) + \frac{1}{\sin^2\phi}\frac{\partial^2 u}{\partial \theta^2}\right].$$

Remark on Notation. Equation (6) is used in calculus and extends the familiar notation for polar coordinates. Unfortunately, some books use θ and ϕ interchanged, an extension of the notation $x = r\cos\phi$, $y = r\sin\phi$ for polar coordinates (used in some European countries).

Boundary Value Problem in Spherical Coordinates

We shall solve the following **Dirichlet problem** in spherical coordinates:

(8)
$$\nabla^2 u = \frac{1}{r^2}\left[\frac{\partial}{\partial r}\left(r^2\frac{\partial u}{\partial r}\right) + \frac{1}{\sin\phi}\frac{\partial}{\partial \phi}\left(\sin\phi\,\frac{\partial u}{\partial \phi}\right)\right] = 0.$$

(9) $$u(R, \phi) = f(\phi)$$

(10) $$\lim_{r\to\infty} u(r, \phi) = 0.$$

The PDE (8) follows from (7) or (7′) by assuming that the solution u will not depend on θ because the Dirichlet condition (9) is independent of θ. This may be an electrostatic potential (or a temperature) $f(\phi)$ at which the sphere S: $r = R$ is kept. Condition (10) means that the potential at infinity will be zero.

Separating Variables by substituting $u(r, \phi) = G(r)H(\phi)$ into (8). Multiplying (8) by r^2, making the substitution and then dividing by GH, we obtain

$$\frac{1}{G}\frac{d}{dr}\left(r^2\frac{dG}{dr}\right) = -\frac{1}{H\sin\phi}\frac{d}{d\phi}\left(\sin\phi\frac{dH}{d\phi}\right).$$

By the usual argument both sides must be equal to a constant k. Thus we get the two ODEs

$$\frac{1}{G}\frac{d}{dr}\left(r^2\frac{dG}{dr}\right) = k \qquad \text{or} \qquad r^2\frac{d^2G}{dr^2} + 2r\frac{dG}{dr} = kG \tag{11}$$

and

$$\frac{1}{\sin\phi}\frac{d}{d\phi}\left(\sin\phi\frac{dH}{d\phi}\right) + kH = 0. \tag{12}$$

The solutions of (11) will take a simple form if we set $k = n(n + 1)$. Then, writing $G' = dG/dr$, etc., we obtain

$$r^2G'' + 2rG' - n(n + 1)\,G = 0. \tag{13}$$

This is an **Euler–Cauchy equation**. From Sec. 2.5 we know that it has solutions $G = r^a$. Substituting this and dropping the common factor r^a gives

$$a(a - 1) + 2a - n(n + 1) = 0. \qquad \text{The roots are} \qquad a = n \quad \text{and} \quad -n - 1.$$

Hence solutions are

$$G_n(r) = r^n \qquad \text{and} \qquad G_n^*(r) = \frac{1}{r^{n+1}}. \tag{14}$$

We now solve (12). Setting $\cos\phi = w$, we have $\sin^2\phi = 1 - w^2$ and

$$\frac{d}{d\phi} = \frac{d}{dw}\frac{dw}{d\phi} = -\sin\phi\,\frac{d}{dw}.$$

Consequently, (12) with $k = n(n + 1)$ takes the form

$$\frac{d}{dw}\left[(1 - w^2)\frac{dH}{dw}\right] + n(n + 1)H = 0. \tag{15}$$

This is **Legendre's equation** (see Sec. 5.3), written out

(15′)
$$(1 - w^2)\frac{d^2H}{dw^2} - 2w\frac{dH}{dw} + n(n + 1)H = 0.$$

For integer $n = 0, 1, \cdots$ the Legendre polynomials

$$H = P_n(w) = P_n(\cos\phi) \qquad n = 0, 1, \cdots,$$

are solutions of Legendre's equation (15). We thus obtain the following two sequences of solution $u = GH$ of Laplace's equation (8), with constant A_n and B_n, where $n = 0, 1, \cdots,$

(16)
$$\text{(a)} \quad u_n(r, \phi) = A_n r^n P_n(\cos\phi), \qquad \text{(b)} \quad u_n^*(r, \phi) = \frac{B_n}{r^{n+1}} P_n(\cos\phi)$$

Use of Fourier–Legendre Series

Interior Problem: Potential Within the Sphere *S*. We consider a series of terms from (16a),

(17)
$$u(r, \phi) = \sum_{n=0}^{\infty} A_n r^n P_n(\cos\phi) \qquad (r \leqq R).$$

Since S is given by $r = R$, for (17) to satisfy the Dirichlet condition (9) on the sphere S, we must have

(18)
$$u(R, \phi) = \sum_{n=0}^{\infty} A_n R^n P_n(\cos\phi) = f(\phi);$$

that is, (18) must be the **Fourier–Legendre series** of $f(\phi)$. From (7) in Sec. 5.8 we get the coefficients

(19*)
$$A_n R^n = \frac{2n + 1}{2}\int_{-1}^{1} \tilde{f}(w)\, P_n(w)\, dw$$

where $\tilde{f}(w)$ denotes $f(\phi)$ as a function of $w = \cos\phi$. Since $dw = -\sin\phi\, d\phi$, and the limits of integration -1 and 1 correspond to $\phi = \pi$ and $\phi = 0$, respectively, we also obtain

(19)
$$A_n = \frac{2n + 1}{2R^n}\int_0^{\pi} f(\phi) P_n(\cos\phi)\sin\phi\, d\phi, \qquad n = 0, 1, \cdots.$$

If $f(\phi)$ and $f'(\phi)$ are piecewise continuous on the interval $0 \leqq \phi \leqq \pi$, then the series (17) with coefficients (19) solves our problem for points inside the sphere because it can be shown that under these continuity assumptions the series (17) with coefficients (19) gives the derivatives occurring in (8) by termwise differentiation, thus justifying our derivation.

Exterior Problem: Potential Outside the Sphere *S*. Outside the sphere we cannot use the functions u_n in (16a) because they do not satisfy (10). But we can use the u_n^* in (16b), which do satisfy (10) (but could not be used inside *S;* why?). Proceeding as before leads to the solution of the exterior problem

$$u(r, \phi) = \sum_{n=0}^{\infty} \frac{B_n}{r^{n+1}} P_n(\cos \phi) \qquad (r \geqq R) \tag{20}$$

satisfying (8), (9), (10), with coefficients

$$B_n = \frac{2n+1}{2} R^{n+1} \int_0^{\pi} f(\phi) P_n(\cos \phi) \sin \phi \, d\phi. \tag{21}$$

The next example illustrates all this for a sphere of radius 1 consisting of two hemispheres that are separated by a small strip of insulating material along the equator, so that these hemispheres can be kept at different potentials (110 V and 0 V).

EXAMPLE 1 **Spherical Capacitor**

Find the potential inside and outside a spherical capacitor consisting of two metallic hemispheres of radius 1 ft separated by a small slit for reasons of insulation, if the upper hemisphere is kept at 110 V and the lower is grounded (Fig. 313).

Solution. The given boundary condition is (recall Fig. 312)

$$f(\phi) = \begin{cases} 110 & \text{if} \quad 0 \leqq \phi < \pi/2 \\ 0 & \text{if} \quad \pi/2 < \phi \leqq \pi. \end{cases}$$

Since $R = 1$, we thus obtain from (19)

$$A_n = \frac{2n+1}{2} \cdot 110 \int_0^{\pi/2} P_n(\cos \phi) \sin \phi \, d\phi$$

$$= \frac{2n+1}{2} \cdot 110 \int_0^{1} P_n(w) \, dw$$

where $w = \cos \phi$. Hence $P_n(\cos \phi) \sin \phi \, d\phi = -P_n(w) \, dw$, we integrate from 1 to 0, and we finally get rid of the minus by integrating from 0 to 1. You can evaluate this integral by your CAS or continue by using (11) in Sec. 5.2, obtaining

$$A_n = 55(2n+1) \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{2^n m!(n-m)!(n-2m)!} \int_0^1 w^{n-2m} \, dw$$

where $M = n/2$ for even n and $M = (n-1)/2$ for odd n. The integral equals $1/(n-2m+1)$. Thus

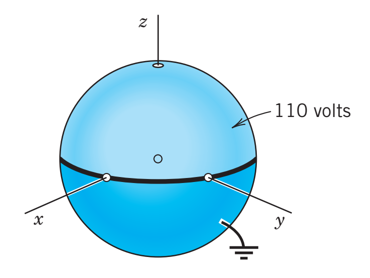


Fig. 313. Spherical capacitor in Example 1

$$A_n = \frac{55(2n+1)}{2^n} \sum_{m=0}^{M} (-1)^m \frac{(2n-2m)!}{m!(n-m)!(n-2m+1)!}. \tag{22}$$

Taking $n = 0$, we get $A_0 = 55$ (since $0! = 1$). For $n = 1, 2, 3, \cdots$ we get

$$A_1 = \frac{165}{2} \cdot \frac{2!}{0!1!2!} = \frac{165}{2},$$

$$A_2 = \frac{275}{4} \left(\frac{4!}{0!2!3!} - \frac{2!}{1!1!1!} \right) = 0,$$

$$A_3 = \frac{385}{8} \left(\frac{6!}{0!3!4!} - \frac{4!}{1!2!2!} \right) = -\frac{385}{8}, \qquad \text{etc.}$$

Hence the ***potential* (17) *inside the sphere*** is (since $P_0 = 1$)

$$u(r, \phi) = 55 + \frac{165}{2} r P_1(\cos\phi) - \frac{385}{8} r^3 P_3(\cos\phi) + \cdots \qquad \text{(Fig. 314)} \tag{23}$$

with $P_1, P_3, \cdots$ given by (11′), Sec. 5.21. Since $R = 1$, we see from (19) and (21) in this section that $B_n = A_n$, and (20) thus gives the ***potential outside the sphere***

$$u(r, \phi) = \frac{55}{r} + \frac{165}{2r^2} P_1(\cos\phi) - \frac{385}{8r^4} P_3(\cos\phi) + \cdots. \tag{24}$$

Partial sums of these series can now be used for computing approximate values of the inner and outer potential. Also, it is interesting to see that far away from the sphere the potential is approximately that of a point charge, namely, $55/r$. (Compare with Theorem 3 in Sec. 9.7.) ■

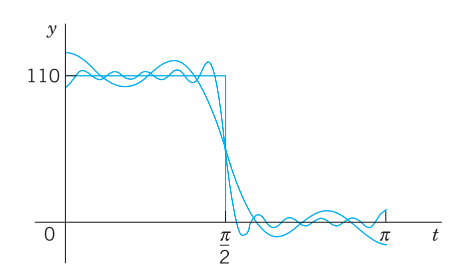


Fig. 314. Partial sums of the first 4, 6, and 11 nonzero terms of (23) for $r = R = 1$

EXAMPLE 2 Simpler Cases. Help with Problems

The technicalities encountered in cases that are similar to the one shown in Example 1 can often be avoided. For instance, find the potential inside the sphere S: $r = R = 1$ when S is kept at the potential $f(\phi) = \cos 2\phi$. (Can you see the potential on S? What is it at the North Pole? The equator? The South Pole?)

Solution. $w = \cos\phi$, $\cos 2\phi = 2\cos^2\phi - 1 = 2w^2 - 1 = \frac{4}{3}P_2(w) - \frac{1}{3} = \frac{4}{3}(\frac{3}{2}w^2 - \frac{1}{2}) - \frac{1}{3}$. Hence the potential in the interior of the sphere is

$$u = \tfrac{4}{3}r^2P_2(w) - \tfrac{1}{3} = \tfrac{4}{3}r^2P_2(\cos\phi) - \tfrac{1}{3} = \tfrac{2}{3}r^2(3\cos^2\phi - 1) - \tfrac{1}{3}.$$ ■

PROBLEM SET 12.11

1. **Spherical coordinates.** Derive (7) from $\nabla^2 u$ in spherical coordinates.
2. **Cylindrical coordinates.** Verify (5) by transforming $\nabla^2 u$ back into Cartesian coordinates.
3. Sketch $P_n(\cos\theta)$, $0 \leqq \theta \leqq 2\pi$, for $n = 0, 1, 2$. (Use (11′) in Sec. 5.2.)
4. **Zero surfaces.** Find the surfaces on which u_1, u_2, u_3 in (16) are **zero**.

5. CAS PROBLEM. Partial Sums. In Example 1 in the text verify the values of A_0, A_1, A_2, A_3 and compute $A_4, \cdots, A_{10}$. Try to find out graphically how well the corresponding partial sums of (23) approximate the given boundary function.

6. CAS EXPERIMENT. Gibbs Phenomenon. Study the Gibbs phenomenon in Example 1 (Fig. 314) graphically.

7. Verify that u_n and u_n^* in (16) are solutions of (8).

8–15 POTENTIALS DEPENDING ONLY ON r

8. Dimension 3. Verify that the potential $u = c/r$, $r = \sqrt{x^2 + y^2 + z^2}$ satisfies Laplace's equation in spherical coordinates.

9. Spherical symmetry. Show that the only solution of Laplace's equation depending only on $r = \sqrt{x^2 + y^2 + z^2}$ is $u = c/r + k$ with constant c and k.

10. Cylindrical symmetry. Show that the only solution of Laplace's equation depending only on $r = \sqrt{x^2 + y^2}$ is $u = c \ln r + k$.

11. Verification. Substituting $u(r)$ with r as in Prob. 9 into $u_{xx} + u_{yy} + u_{zz} = 0$, verify that $u'' + 2u'/r = 0$, in agreement with (7).

12. Dirichlet problem. Find the electrostatic potential between coaxial cylinders of radii $r_1 = 2$ cm and $r_2 = 4$ cm kept at the potentials $U_1 = 220$ V and $U_2 = 140$ V, respectively.

13. Dirichlet problem. Find the electrostatic potential between two concentric spheres of radii $r_1 = 2$ cm and $r_2 = 4$ cm kept at the potentials $U_1 = 220$ V and $U_2 = 140$ V, respectively. Sketch and compare the equipotential lines in Probs. 12 and 13. Comment.

14. Heat problem. If the surface of the ball $r^2 = x^2 + y^2 + z^2 \leqq R^2$ is kept at temperature zero and the initial temperature in the ball is $f(r)$, show that the temperature $u(r, t)$ in the ball is a solution of $u_t = c^2(u_{rr} + 2u_r/r)$ satisfying the conditions $u(R, t) = 0$, $u(r, 0) = f(r)$. Show that setting $v = ru$ gives $v_t = c^2 v_{rr}$, $v(R, t) = 0$, $v(r, 0) = rf(r)$. Include the condition $v(0, t) = 0$ (which holds because u must be bounded at $r = 0$), and solve the resulting problem by separating variables.

15. What are the analogs of Probs. 12 and 13 in heat conduction?

16–20 BOUNDARY VALUE PROBLEMS IN SPHERICAL COORDINATES r, θ, ϕ

Find the potential in the interior of the sphere $r = R = 1$ if the interior is free of charges and the potential on the sphere is

16. $f(\phi) = \cos \phi$

17. $f(\phi) = 1$

18. $f(\phi) = 1 - \cos^2 \phi$

19. $f(\phi) = \cos 2\phi$

20. $f(\phi) = 10 \cos^3 \phi - 3 \cos^2 \phi - 5 \cos \phi - 1$

21. Point charge. Show that in Prob. 17 the potential exterior to the sphere is the same as that of a point charge at the origin.

22. Exterior potential. Find the potentials exterior to the sphere in Probs. 16 and 19.

23. Plane intersections. Sketch the intersections of the equipotential surfaces in Prob. 16 with xz-plane.

24. TEAM PROJECT. Transmission Line and Related PDEs. Consider a long cable or telephone wire (Fig. 315) that is imperfectly insulated, so that leaks occur along the entire length of the cable. The source S of the current $i(x, t)$ in the cable is at $x = 0$, the receiving end T at $x = l$. The current flows from S to T and through the load, and returns to the ground. Let the constants R, L, C, and G denote the resistance, inductance, capacitance to ground, and conductance to ground, respectively, of the cable per unit length.

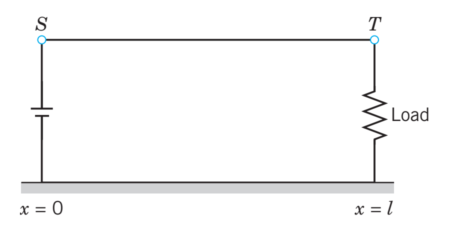


Fig. 315. Transmission line

(a) Show that (**"first transmission line equation"**)

$$-\frac{\partial u}{\partial x} = Ri + L\frac{\partial i}{\partial t}$$

where $u(x, t)$ is the potential in the cable. *Hint:* Apply Kirchhoff's voltage law to a small portion of the cable between x and $x + \Delta x$ (difference of the potentials at x and $x + \Delta x$ = resistive drop + inductive drop).

(b) Show that for the cable in (a) (**"second transmission line equation"**),

$$-\frac{\partial i}{\partial x} = Gu + C\frac{\partial u}{\partial t}.$$

Hint: Use Kirchhoff's current law (difference of the currents at x and $x + \Delta x$ = loss due to leakage to ground + capacitive loss).

(c) Second-order PDEs. Show that elimination of i or u from the transmission line equations leads to

$$u_{xx} = LCu_{tt} + (RC + GL)u_t + RGu,$$
$$i_{xx} = LCi_{tt} + (RC + GL)i_t + RGi.$$

(d) Telegraph equations. For a submarine cable, G is negligible and the frequencies are low. Show that this leads to the so-called *submarine cable equations* or **telegraph equations**

$$u_{xx} = RCu_t, \qquad i_{xx} = RCi_t.$$

Find the potential in a submarine cable with ends ($x = 0, x = l$) grounded and initial voltage distribution $U_0 = \text{const}$.

(e) High-frequency line equations. Show that in the case of alternating currents of high frequencies the equations in (c) can be approximated by the so-called **high-frequency line equations**

$$u_{xx} = LCu_{tt}, \qquad i_{xx} = LCi_{tt}.$$

Solve the first of them, assuming that the initial potential is

$$U_0 \sin(\pi x/l),$$

and $u_t(x, 0) = 0$ and $u = 0$ at the ends $x = 0$ and $x = l$ for all t.

25. Reflection in a sphere. Let r, θ, ϕ be spherical coordinates. If $u(r, \theta, \phi)$ satisfies $\nabla^2 u = 0$, show that $v(r, \theta, \phi) = u(1/r, \theta, \phi)/r$ satisfies $\nabla^2 v = 0$.

12.12 Solution of PDEs by Laplace Transforms

Readers familiar with Chap. 6 may wonder whether Laplace transforms can also be used for solving *partial* differential equations. The answer is yes, particularly if one of the independent variables ranges over the positive axis. The steps to obtain a solution are similar to those in Chap. 6. For a PDE in two variables they are as follows.

1. Take the Laplace transform with respect to one of the two variables, usually t. This gives an ***ODE for the transform*** of the unknown function. This is so since the derivatives of this function with respect to the other variable slip into the transformed equation. The latter also incorporates the given boundary and initial conditions.
2. Solving that ODE, obtain the transform of the unknown function.
3. Taking the inverse transform, obtain the solution of the given problem.

If the coefficients of the given equation do not depend on t, the use of Laplace transforms will simplify the problem.

We explain the method in terms of a typical example.

EXAMPLE 1 Semi-Infinite String

Find the displacement $w(x, t)$ of an elastic string subject to the following conditions. (We write w since we need u to denote the unit step function.)

(i) The string is initially at rest on the x-axis from $x = 0$ to ∞ (*"semi-infinite string"*).

(ii) For $t > 0$ the left end of the string ($x = 0$) is moved in a given fashion, namely, according to a single sine wave

$$w(0, t) = f(t) = \begin{cases} \sin t & \text{if } 0 \leqq t \leqq 2\pi \\ 0 & \text{otherwise} \end{cases} \qquad \text{(Fig. 316).}$$

(iii) Furthermore, $\lim_{x \to \infty} w(x, t) = 0$ for $t \geqq 0$.

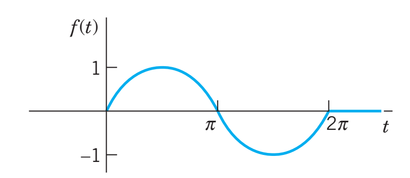


Fig. 316. Motion of the left end of the string in Example 1 as a function of time t

Of course there is no infinite string, but our model describes a long string or rope (of negligible weight) with its right end fixed far out on the x-axis.

Solution. We have to solve the wave equation (Sec. 12.2)

$$\frac{\partial^2 w}{\partial t^2} = c^2 \frac{\partial^2 w}{\partial x^2}, \qquad c^2 = \frac{T}{\rho} \tag{1}$$

for positive x and t, subject to the "boundary conditions"

$$w(0, t) = f(t), \qquad \lim_{x \to \infty} w(x, t) = 0 \qquad (t \geqq 0) \tag{2}$$

with f as given above, and the initial conditions

$$\text{(a)} \quad w(x, 0) = 0, \qquad \text{(b)} \quad w_t(x, 0) = 0. \tag{3}$$

We take the Laplace transform ***with respect to t***. By (2) in Sec. 6.2,

$$\mathcal{L}\left\{\frac{\partial^2 w}{\partial t^2}\right\} = s^2\mathcal{L}\{w\} - sw(x, 0) - w_t(x, 0) = c^2\mathcal{L}\left\{\frac{\partial^2 w}{\partial x^2}\right\}.$$

The expression $-sw(x, 0) - w_t(x, 0)$ drops out because of (3). On the right we assume that we may interchange integration and differentiation. Then

$$\mathcal{L}\left\{\frac{\partial^2 w}{\partial x^2}\right\} = \int_0^\infty e^{-st}\frac{\partial^2 w}{\partial x^2}\,dt = \frac{\partial^2}{\partial x^2}\int_0^\infty e^{-st}w(x, t)\,dt = \frac{\partial^2}{\partial x^2}\mathcal{L}\{w(x, t)\}.$$

Writing $W(x, s) = \mathcal{L}\{w(x, t)\}$, we thus obtain

$$s^2 W = c^2\frac{\partial^2 W}{\partial x^2}, \qquad \text{thus} \qquad \frac{\partial^2 W}{\partial x^2} - \frac{s^2}{c^2}W = 0.$$

Since this equation contains only a derivative with respect to x, it may be regarded as an ***ordinary differential equation*** for $W(x, s)$ considered as a function of x. A general solution is

$$W(x, s) = A(s)e^{sx/c} + B(s)e^{-sx/c}. \tag{4}$$

From (2) we obtain, writing $F(s) = \mathcal{L}\{f(t)\}$,

$$W(0, s) = \mathcal{L}\{w(0, t)\} = \mathcal{L}\{f(t)\} = F(s).$$

Assuming that we can interchange integration and taking the limit, we have

$$\lim_{x\to\infty} W(x, s) = \lim_{x\to\infty}\int_0^\infty e^{-st}w(x, t)\,dt = \int_0^\infty e^{-st}\lim_{x\to\infty} w(x, t)\,dt = 0.$$

This implies $A(s) = 0$ in (4) because $c > 0$, so that for every fixed positive s the function $e^{sx/c}$ increases as x increases. Note that we may assume $s > 0$ since a Laplace transform generally exists for *all* s greater than some fixed k (Sec. 6.2). Hence we have

$$W(0, s) = B(s) = F(s),$$

so that (4) becomes

$$W(x, s) = F(s)e^{-sx/c}.$$

From the second shifting theorem (Sec. 6.3) with $a = x/c$ we obtain the inverse transform

$$w(x, t) = f\left(t - \frac{x}{c}\right)u\left(t - \frac{x}{c}\right) \qquad \text{(Fig. 317)} \tag{5}$$

that is,

$$w(x, t) = \sin\left(t - \frac{x}{c}\right) \quad \text{if} \quad \frac{x}{c} < t < \frac{x}{c} + 2\pi \quad \text{or} \quad ct > x > (t - 2\pi)c$$

and zero otherwise. This is a single sine wave traveling to the right with speed c. Note that a point x remains at rest until $t = x/c$, the time needed to reach that x if one starts at $t = 0$ (start of the motion of the left end) and travels with speed c. The result agrees with our physical intuition. Since we proceeded formally, we must verify that (5) satisfies the given conditions. We leave this to the student. ■

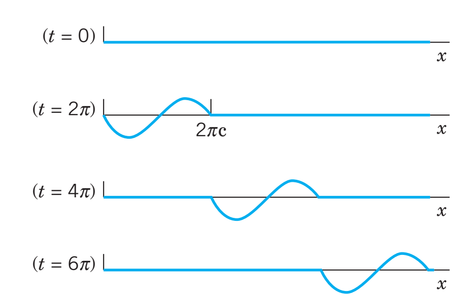


Fig. 317. Traveling wave in Example 1

We have reached the end of Chapter 12, in which we concentrated on the most important partial differential equations (PDEs) in physics and engineering. We have also reached the end of Part C on Fourier Analysis and PDEs.

Outlook

We have seen that PDEs underlie the modeling process of various important engineering application. Indeed, PDEs are the subject of many ongoing research projects.

Numerics for PDEs follows in Secs. 21.4–21.7, which, by design for greater flexibility in teaching, are independent of the other sections in Part E on numerics.

In the next part, that is, Part D on **complex analysis**, we turn to an area of a different nature that is also highly important to the engineer. The rich vein of examples and problems will signify this. It is of note that Part D includes another approach to the two-dimensional **Laplace equation** with applications, as shown in Chap. 18.

PROBLEM SET 12.12

1. Verify the solution in Example 1. What traveling wave do we obtain in Example 1 for a nonterminating sinusoidal motion of the left end starting at $t = 2\pi$?

2. Sketch a figure similar to Fig. 317 when $c = 1$ and $f(x)$ is "triangular," say, $f(x) = x$ if $0 < x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} < x < 1$ and 0 otherwise.

3. How does the speed of the wave in Example 1 of the text depend on the tension and on the mass of the string?

4–8 SOLVE BY LAPLACE TRANSFORMS

4. $\frac{\partial w}{\partial x} + x\frac{\partial w}{\partial t} = x, \quad w(x, 0) = 1, \quad w(0, t) = 1$

5. $x\frac{\partial w}{\partial x} + \frac{\partial w}{\partial t} = xt, \quad w(x, 0) = 0 \text{ if } x \geqq 0, \quad w(0, t) = 0 \text{ if } t \geqq 0$

6. $\frac{\partial w}{\partial x} + 2x\frac{\partial w}{\partial t} = 2x, \quad w(x, 0) = 1, \quad w(0, t) = 1$

7. Solve Prob. 5 by separating variables.

8. $\frac{\partial^2 w}{\partial x^2} = 100\frac{\partial^2 w}{\partial t^2} + 100\frac{\partial w}{\partial t} + 25w,$

$w(x, 0) = 0$ if $x \geqq 0$, $\quad w_t(x, 0) = 0$ if $t \geqq 0$,
$w(0, t) = \sin t$ if $t \geqq 0$

9–12 HEAT PROBLEM

Find the temperature $w(x, t)$ in a semi-infinite laterally insulated bar extending from $x = 0$ along the x-axis to infinity, assuming that the initial temperature is 0, $w(x, t) \to 0$ as $x \to \infty$ for every fixed $t \geqq 0$, and $w(0, t) = f(t)$. Proceed as follows.

9. Set up the model and show that the Laplace transform leads to

$$sW = c^2 \frac{\partial^2 W}{\partial x^2} \quad (W = \mathcal{L}\{w\})$$

and

$$W = F(s)e^{-\sqrt{s}x/c} \quad (F = \mathcal{L}\{f\}).$$

10. Applying the convolution theorem, show that in Prob. 9,

$$w(x, t) = \frac{x}{2c\sqrt{\pi}} \int_0^t f(t - \tau)\tau^{-3/2} e^{-x^2/(4c^2\tau)} d\tau.$$

11. Let $w(0, t) = f(t) = u(t)$ (Sec. 6.3). Denote the corresponding w, W, and F by w_0, W_0, and F_0. Show that then in Prob. 10,

$$w_0(x, t) = \frac{x}{2c\sqrt{\pi}} \int_0^t \tau^{-3/2} e^{-x^2/(4c^2\tau)} d\tau$$
$$= 1 - \operatorname{erf}\left(\frac{x}{2c\sqrt{t}}\right)$$

with the error function erf as defined in Problem Set 12.7.

12. **Duhamel's formula.**[4] Show that in Prob. 11,

$$W_0(x, s) = \frac{1}{s} e^{-\sqrt{s}x/c}$$

and the convolution theorem gives *Duhamel's formula*

$$W(x, t) = \int_0^t f(t - \tau) \frac{\partial w_0}{\partial \tau} d\tau.$$

CHAPTER 12 REVIEW QUESTIONS AND PROBLEMS

1. For what kinds of problems will modeling lead to an ODE? To a PDE?
2. Mention some of the basic physical principles or laws that will give a PDE in modeling.
3. State three or four of the most important PDEs and their main applications.
4. What is "separating variables" in a PDE? When did we apply it twice in succession?
5. What is d'Alembert's solution method? To what PDE does it apply?
6. What role did Fourier series play in this chapter? Fourier integrals?
7. When and why did Legendre's equation occur? Bessel's equation?
8. What are the eigenfunctions and their frequencies of the vibrating string? Of the vibrating membrane?
9. What do you remember about types of PDEs? Normal forms? Why is this important?
10. When did we use polar coordinates? Cylindrical coordinates? Spherical coordinates?
11. Explain mathematically (not physically) why we got exponential functions in separating the heat equation, but not for the wave equation.
12. Why and where did the error function occur?
13. How do problems for the wave equation and the heat equation differ regarding additional conditions?
14. Name and explain the three kinds of boundary conditions for Laplace's equation.
15. Explain how the Laplace transform applies to PDEs.

16–18 Solve for $u = u(x, y)$:

16. $u_{xx} + 25u = 0$
17. $u_{yy} + u_y - 6u = 18$
18. $u_{xx} + u_x = 0, \quad u(0, y) = f(y), \quad u_x(0, y) = g(y)$

19–21 NORMAL FORM

Transform to normal form and solve:

19. $u_{xy} = u_{yy}$
20. $u_{xx} + 6u_{xy} + 9u_{yy} = 0$
21. $u_{xx} - 4u_{yy} = 0$

22–24 VIBRATING STRING

Find and sketch or graph (as in Fig. 288 in Sec. 12.3) the deflection $u(x, t)$ of a vibrating string of length π, extending from $x = 0$ to $x = \pi$, and $c^2 = T/\rho = 4$ starting with velocity zero and deflection:

22. $\sin 4x$
23. $\sin^3 x$
24. $\frac{1}{2}\pi - |x - \frac{1}{2}\pi|$

[4] JEAN–MARIE CONSTANT DUHAMEL (1797–1872), French mathematician.

25–27 HEAT

Find the temperature distribution in a laterally insulated thin copper bar ($c^2 = K/(\sigma\rho) = 1.158 \text{ cm}^2/\text{sec}$) of length 100 cm and constant cross section with endpoints at $x = 0$ and 100 kept at 0°C and initial temperature:

25. $\sin 0.01\pi x$

26. $50 - |50 - x|$

27. $\sin^3 0.01\pi x$

28–30 ADIABATIC CONDITIONS

Find the temperature distribution in a laterally insulated bar of length π with $c^2 = 1$ for the adiabatic boundary condition (see Problem Set 12.6) and initial temperature:

28. $3x^2$

29. $100 \cos 2x$

30. $2\pi - 4|x - \frac{1}{2}\pi|$

31–32 TEMPERATURE IN A PLATE

31. Let $f(x, y) = u(x, y, 0)$ be the initial temperature in a thin square plate of side π with edges kept at 0°C and faces perfectly insulated. Separating variables, obtain from $u_t = c^2\nabla^2 u$ the solution

$$u(x, y, t) = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} B_{mn} \sin mx \sin ny \, e^{-c^2(m^2+n^2)t}$$

where

$$B_{mn} = \frac{4}{\pi^2}\int_0^{\pi}\int_0^{\pi} f(x, y) \sin mx \sin ny \, dx \, dy.$$

32. Find the temperature in Prob. 31 if $f(x, y) = x(\pi - x)y(\pi - y)$.

33–37 MEMBRANES

Show that the following membranes of area 1 with $c^2 = 1$ have the frequencies of the fundamental mode as given (4-decimal values). Compare.

33. Circle: $\alpha_1/(2\sqrt{\pi}) = 0.6784$

34. Square: $1/\sqrt{2} = 0.7071$

35. Rectangle with sides 1:2: $\sqrt{5/8} = 0.7906$

36. Semicircle: $3.832/\sqrt{8\pi} = 0.7643$

37. **Quadrant** of circle: $\alpha_{21}/(4\sqrt{\pi}) = 0.7244$ ($\alpha_{21} = 5.13562$ = first positive zero of J_2)

38–40 ELECTROSTATIC POTENTIAL

Find the potential in the following charge-free regions.

38. Between two concentric spheres of radii r_0 and r_1 kept at potentials u_0 and u_1, respectively.

39. Between two coaxial circular cylinders of radii r_0 and r_1 kept at the potentials u_0 and u_1, respectively. Compare with Prob. 38.

40. In the interior of a sphere of radius 1 kept at the potential $f(\phi) = \cos 3\phi + 3 \cos \phi$ (referred to our usual spherical coordinates).

SUMMARY OF CHAPTER 12
Partial Differential Equations (PDEs)

Whereas ODEs (Chaps. 1–6) serve as models of problems involving only *one* independent variable, problems involving *two or more* independent variables (space variables or time t and one or several space variables) lead to PDEs. This accounts for the enormous importance of PDEs to the engineer and physicist. Most important are:

(1) $u_{tt} = c^2 u_{xx}$ One-dimensional wave equation (Secs. 12.2–12.4)

(2) $u_{tt} = c^2(u_{xx} + u_{yy})$ Two-dimensional wave equation (Secs. 12.8–12.10)

(3) $u_t = c^2 u_{xx}$ One-dimensional heat equation (Secs. 12.5, 12.6, 12.7)

(4) $\nabla^2 u = u_{xx} + u_{yy} = 0$ Two-dimensional Laplace equation (Secs. 12.6, 12.10)

(5) $\nabla^2 u = u_{xx} + u_{yy} + u_{zz} = 0$ Three-dimensional Laplace equation (Sec. 12.11).

Equations (1) and (2) are hyperbolic, (3) is parabolic, (4) and (5) are elliptic.

In practice, one is interested in obtaining the solution of such an equation in a given region satisfying given additional conditions, such as **initial conditions** (conditions at time $t = 0$) or **boundary conditions** (prescribed values of the solution u or some of its derivatives on the boundary surface S, or boundary curve C, of the region) or both. For (1) and (2) one prescribes two initial conditions (initial displacement and initial velocity). For (3) one prescribes the initial temperature distribution. For (4) and (5) one prescribes a boundary condition and calls the resulting problem a (see Sec. 12.6)

Dirichlet problem if u is prescribed on S,
Neumann problem if $u_n = \partial u/\partial n$ is prescribed on S,
Mixed problem if u is prescribed on one part of S and u_n on the other.

A general method for solving such problems is the method of **separating variables** or **product method**, in which one assumes solutions in the form of products of functions each depending on one variable only. Thus equation (1) is solved by setting $u(x, t) = F(x)G(t)$; see Sec. 12.3; similarly for (3) (see Sec. 12.6). Substitution into the given equation yields ***ordinary*** differential equations for F and G, and from these one gets infinitely many solutions $F = F_n$ and $G = G_n$ such that the corresponding functions

$$u_n(x, t) = F_n(x)G_n(t)$$

are solutions of the PDE satisfying the given boundary conditions. These are the **eigenfunctions** of the problem, and the corresponding **eigenvalues** determine the frequency of the vibration (or the rapidity of the decrease of temperature in the case of the heat equation, etc.). To satisfy also the initial condition (or conditions), one must consider infinite series of the u_n, whose coefficients turn out to be the Fourier coefficients of the functions f and g representing the given initial conditions (Secs. 12.3, 12.6). Hence **Fourier series** (and *Fourier integrals*) are of basic importance here (Secs. 12.3, 12.6, 12.7, 12.9).

Steady-state problems are problems in which the solution does not depend on time t. For these, the heat equation $u_t = c^2\nabla^2 u$ becomes the Laplace equation.

Before solving an initial or boundary value problem, one often transforms the PDE into coordinates in which the boundary of the region considered is given by simple formulas. Thus in polar coordinates given by $x = r\cos\theta$, $y = r\sin\theta$, the **Laplacian** becomes (Sec. 12.11)

$$\nabla^2 u = u_{rr} + \frac{1}{r}u_r + \frac{1}{r^2}u_{\theta\theta}; \tag{6}$$

for spherical coordinates see Sec. 12.10. If one now separates the variables, one gets ***Bessel's equation*** from (2) and (6) (vibrating circular membrane, Sec. 12.10) and ***Legendre's equation*** from (5) transformed into spherical coordinates (Sec. 12.11).

PART D

Complex Analysis

Complex analysis has many applications in heat conduction, fluid flow, electrostatics, and in other areas. It extends the familiar "real calculus" to "complex calculus" by introducing complex numbers and functions. While many ideas carry over from calculus to complex analysis, there is a marked difference between the two. For example, analytic functions, which are the "good functions" (differentiable in some domain) of complex analysis, have derivatives of all orders. This is in contrast to calculus, where real-valued functions of real variables may have derivatives only up to a certain order. Thus, in certain ways, problems that are difficult to solve in real calculus may be much easier to solve in complex analysis. Complex analysis is important in applied mathematics for three main reasons:

1. Two-dimensional potential problems can be modeled and solved by methods of analytic functions. This reason is the real and imaginary parts of analytic functions satisfy Laplace's equation in two real variables.

2. Many difficult integrals (real or complex) that appear in applications can be solved quite elegantly by complex integration.

3. Most functions in engineering mathematics are analytic functions, and their study as functions of a complex variable leads to a deeper understanding of their properties and to interrelations in complex that have no analog in real calculus.

Complex Numbers and Functions. Complex Differentiation

The transition from "real calculus" to "complex calculus" starts with a discussion of ***complex numbers*** and their geometric representation in the ***complex plane***. We then progress to ***analytic functions*** in Sec. 13.3. We desire functions to be analytic because these are the "useful functions" in the sense that they are differentiable in some domain and operations of complex analysis can be applied to them. The most important equations are therefore the Cauchy–Riemann equations in Sec. 13.4 because they allow a test of analyticity of such functions. Moreover, we show how the Cauchy–Riemann equations are related to the important ***Laplace equation***.

The remaining sections of the chapter are devoted to elementary complex functions (exponential, trigonometric, hyperbolic, and logarithmic functions). These generalize the familiar real functions of calculus. Detailed knowledge of them is an absolute necessity in practical work, just as that of their real counterparts is in calculus.

Prerequisite: Elementary calculus.
References and Answers to Problems: App. 1 Part D, App. 2.

13.1 Complex Numbers and Their Geometric Representation

The material in this section will most likely be familiar to the student and serve as a review.

Equations without *real* solutions, such as $x^2 = -1$ or $x^2 - 10x + 40 = 0$, were observed early in history and led to the introduction of complex numbers.[1] By definition, a **complex number** z is an ordered pair (x, y) of real numbers x and y, written

$$z = (x, y).$$

[1]First to use complex numbers for this purpose was the Italian mathematician GIROLAMO CARDANO (1501–1576), who found the formula for solving cubic equations. The term "complex number" was introduced by CARL FRIEDRICH GAUSS (see the footnote in Sec. 5.4), who also paved the way for a general use of complex numbers.

x is called the **real part** and y the **imaginary part** of z, written

$$x = \text{Re}\, z, \qquad y = \text{Im}\, z.$$

By definition, two complex numbers are **equal** if and only if their real parts are equal and their imaginary parts are equal.

(0, 1) is called the **imaginary unit** and is denoted by i,

(1) $$i = (0, 1).$$

Addition, Multiplication. Notation $z = x + iy$

Addition of two complex numbers $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$ is defined by

(2) $$z_1 + z_2 = (x_1, y_1) + (x_2, y_2) = (x_1 + x_2, \quad y_1 + y_2).$$

Multiplication is defined by

(3) $$z_1 z_2 = (x_1, y_1)(x_2, y_2) = (x_1 x_2 - y_1 y_2, \quad x_1 y_2 + x_2 y_1).$$

These two definitions imply that

$$(x_1, 0) + (x_2, 0) = (x_1 + x_2, 0)$$

and

$$(x_1, 0)(x_2, 0) = (x_1 x_2, 0)$$

as for real numbers x_1, x_2. Hence the complex numbers "***extend***" the real numbers. We can thus write

$$(x, 0) = x. \qquad \text{Similarly,} \qquad (0, y) = iy$$

because by (1), and the definition of multiplication, we have

$$iy = (0, 1)y = (0, 1)(y, 0) = (0 \cdot y - 1 \cdot 0, \quad 0 \cdot 0 + 1 \cdot y) = (0, y).$$

Together we have, by addition, $(x, y) = (x, 0) + (0, y) = x + iy$.

In practice, complex numbers $z = (x, y)$ are written

(4) $$z = x + iy$$

or $z = x + yi$, e.g., $17 + 4i$ (instead of $i4$).

Electrical engineers often write j instead of i because they need i for the current.

If $x = 0$, then $z = iy$ and is called **pure imaginary**. Also, (1) and (3) give

(5) $$i^2 = -1$$

because, by the definition of multiplication, $i^2 = ii = (0, 1)(0, 1) = (-1, 0) = -1$.

For **addition** the standard notation (4) gives [see (2)]

$$(x_1 + iy_1) + (x_2 + iy_2) = (x_1 + x_2) + i(y_1 + y_2).$$

For **multiplication** the standard notation gives the following very simple recipe. Multiply each term by each other term and use $i^2 = -1$ when it occurs [see (3)]:

$$\begin{aligned}(x_1 + iy_1)(x_2 + iy_2) &= x_1x_2 + ix_1y_2 + iy_1x_2 + i^2y_1y_2 \\ &= (x_1x_2 - y_1y_2) + i(x_1y_2 + x_2y_1).\end{aligned}$$

This agrees with (3). And it shows that $x + iy$ is a more practical notation for complex numbers than (x, y).

If you know vectors, you see that (2) is vector addition, whereas the multiplication (3) has no counterpart in the usual vector algebra.

EXAMPLE 1 Real Part, Imaginary Part, Sum and Product of Complex Numbers

Let $z_1 = 8 + 3i$ and $z_2 = 9 - 2i$. Then $\text{Re}\, z_1 = 8$, $\text{Im}\, z_1 = 3$, $\text{Re}\, z_2 = 9$, $\text{Im}\, z_2 = -2$ and

$$z_1 + z_2 = (8 + 3i) + (9 - 2i) = 17 + i,$$

$$z_1z_2 = (8 + 3i)(9 - 2i) = 72 + 6 + i(-16 + 27) = 78 + 11i.$$

■

Subtraction, Division

Subtraction and **division** are defined as the inverse operations of addition and multiplication, respectively. Thus the **difference** $z = z_1 - z_2$ is the complex number z for which $z_1 = z + z_2$. Hence by (2),

$$z_1 - z_2 = (x_1 - x_2) + i(y_1 - y_2). \tag{6}$$

The **quotient** $z = z_1/z_2$ $(z_2 \neq 0)$ is the complex number z for which $z_1 = zz_2$. If we equate the real and the imaginary parts on both sides of this equation, setting $z = x + iy$, we obtain $x_1 = x_2x - y_2y$, $y_1 = y_2x + x_2y$. The solution is

$$z = \frac{z_1}{z_2} = x + iy, \qquad x = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2}, \qquad y = \frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}. \tag{7*}$$

The ***practical rule*** used to get this is by multiplying numerator and denominator of z_1/z_2 by $x_2 - iy_2$ and simplifying:

$$z = \frac{x_1 + iy_1}{x_2 + iy_2} = \frac{(x_1 + iy_1)(x_2 - iy_2)}{(x_2 + iy_2)(x_2 - iy_2)} = \frac{x_1x_2 + y_1y_2}{x_2^2 + y_2^2} + i\,\frac{x_2y_1 - x_1y_2}{x_2^2 + y_2^2}. \tag{7}$$

EXAMPLE 2 Difference and Quotient of Complex Numbers

For $z_1 = 8 + 3i$ and $z_2 = 9 - 2i$ we get $z_1 - z_2 = (8 + 3i) - (9 - 2i) = -1 + 5i$ and

$$\frac{z_1}{z_2} = \frac{8 + 3i}{9 - 2i} = \frac{(8 + 3i)(9 + 2i)}{(9 - 2i)(9 + 2i)} = \frac{66 + 43i}{81 + 4} = \frac{66}{85} + \frac{43}{85}i.$$

Check the division by multiplication to get $8 + 3i$. ■

Complex numbers satisfy the same commutative, associative, and distributive laws as real numbers (see the problem set).

Complex Plane

So far we discussed the algebraic manipulation of complex numbers. Consider the geometric representation of complex numbers, which is of great practical importance. We choose two perpendicular coordinate axes, the horizontal x-axis, called the **real axis**, and the vertical y-axis, called the **imaginary axis**. On both axes we choose the same unit of length (Fig. 318). This is called a **Cartesian coordinate system**.

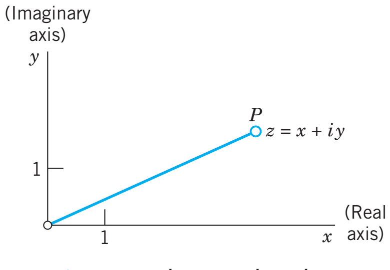


Fig. 318. The complex plane

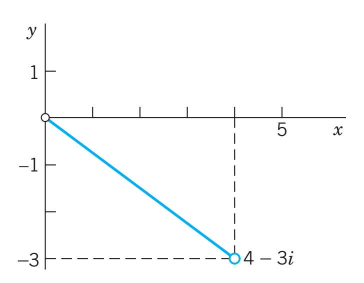


Fig. 319. The number $4 - 3i$ in the complex plane

We now plot a given complex number $z = (x, y) = x + iy$ as the point P with coordinates x, y. The xy-plane in which the complex numbers are represented in this way is called the **complex plane**.[2] Figure 319 shows an example.

Instead of saying "the point represented by z in the complex plane" we say briefly and simply "***the point z in the complex plane***." This will cause no misunderstanding.

Addition and subtraction can now be visualized as illustrated in Figs. 320 and 321.

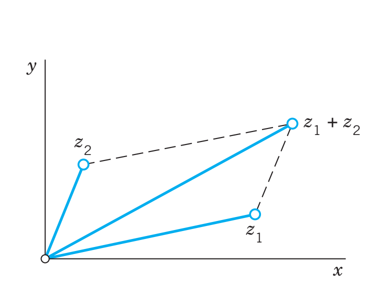


Fig. 320. Addition of complex numbers

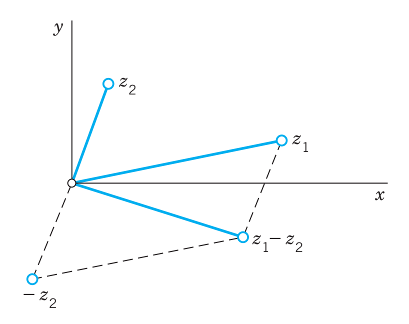


Fig. 321. Subtraction of complex numbers

[2]Sometimes called the **Argand diagram**, after the French mathematician JEAN ROBERT ARGAND (1768–1822), born in Geneva and later librarian in Paris. His paper on the complex plane appeared in 1806, nine years after a similar memoir by the Norwegian mathematician CASPAR WESSEL (1745–1818), a surveyor of the Danish Academy of Science.

Complex Conjugate Numbers

The complex conjugate $\bar{z}$ of a complex number $z = x + iy$ is defined by

$$\bar{z} = x - iy.$$

It is obtained geometrically by reflecting the point z in the real axis. Figure 322 shows this for $z = 5 + 2i$ and its conjugate $\bar{z} = 5 - 2i$.

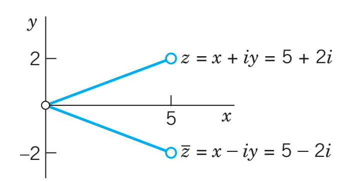


Fig. 322. Complex conjugate numbers

The complex conjugate is important because it permits us to switch from complex to real. Indeed, by multiplication, $z\bar{z} = x^2 + y^2$ (verify!). By addition and subtraction, $z + \bar{z} = 2x$, $z - \bar{z} = 2iy$. We thus obtain for the real part x and the imaginary part y (not iy!) of $z = x + iy$ the important formulas

(8)
$$\text{Re}\, z = x = \tfrac{1}{2}(z + \bar{z}), \qquad \text{Im}\, z = y = \frac{1}{2i}(z - \bar{z}).$$

If z is real, $z = x$, then $\bar{z} = z$ by the definition of $\bar{z}$, and conversely. Working with conjugates is easy, since we have

(9)
$$\overline{(z_1 + z_2)} = \bar{z}_1 + \bar{z}_2, \qquad \overline{(z_1 - z_2)} = \bar{z}_1 - \bar{z}_2,$$
$$\overline{(z_1 z_2)} = \bar{z}_1 \bar{z}_2, \qquad \overline{\left(\frac{z_1}{z_2}\right)} = \frac{\bar{z}_1}{\bar{z}_2}.$$

EXAMPLE 3 Illustration of (8) and (9)

Let $z_1 = 4 + 3i$ and $z_2 = 2 + 5i$. Then by (8),

$$\text{Im}\, z_1 = \frac{1}{2i}[(4 + 3i) - (4 - 3i)] = \frac{3i + 3i}{2i} = 3.$$

Also, the multiplication formula in (9) is verified by

$$\overline{(z_1 z_2)} = \overline{(4 + 3i)(2 + 5i)} = \overline{(-7 + 26i)} = -7 - 26i,$$
$$\bar{z}_1 \bar{z}_2 = (4 - 3i)(2 - 5i) = -7 - 26i.$$
■

PROBLEM SET 13.1

1. **Powers of i.** Show that $i^2 = -1, i^3 = -i, i^4 = 1, i^5 = i, \cdots$ and $1/i = -i, 1/i^2 = -1, 1/i^3 = i, \cdots$.
2. **Rotation.** Multiplication by i is geometrically a counterclockwise rotation through $\pi/2$ (90°). Verify this by graphing z and iz and the angle of rotation for $z = 1 + i, z = -1 + 2i, z = 4 - 3i$.
3. **Division.** Verify the calculation in (7). Apply (7) to $(26 - 18i)/(6 - 2i)$.

4. Law for conjugates. Verify (9) for $z_1 = -11 + 10i$, $z_2 = -1 + 4i$.

5. Pure imaginary number. Show that $z = x + iy$ is pure imaginary if and only if $\bar{z} = -z$.

6. Multiplication. If the product of two complex numbers is zero, show that at least one factor must be zero.

7. Laws of addition and multiplication. Derive the following laws for complex numbers from the corresponding laws for real numbers.

$z_1 + z_2 = z_2 + z_1$, $z_1 z_2 = z_2 z_1$ (*Commutative laws*)

$(z_1 + z_2) + z_3 = z_1 + (z_2 + z_3)$, (*Associative laws*)

$(z_1 z_2) z_3 = z_1 (z_2 z_3)$

$z_1(z_2 + z_3) = z_1 z_2 + z_1 z_3$ (*Distributive law*)

$0 + z = z + 0 = z,$

$z + (-z) = (-z) + z = 0, \qquad z \cdot 1 = z.$

8–15 COMPLEX ARITHMETIC

Let $z_1 = -2 + 11i$, $z_2 = 2 - i$. Showing the details of your work, find, in the form $x + iy$:

8. $z_1 z_2$, $\overline{(z_1 z_2)}$

9. $\text{Re}\,(z_1^2)$, $(\text{Re}\,z_1)^2$

10. $\text{Re}\,(1/z_2^2)$, $1/\text{Re}\,(z_2^2)$

11. $(z_1 - z_2)^2/16$, $(z_1/4 - z_2/4)^2$

12. z_1/z_2, z_2/z_1

13. $(z_1 + z_2)(z_1 - z_2)$, $z_1^2 - z_2^2$

14. $\bar{z}_1/\bar{z}_2$, $\overline{(z_1/z_2)}$

15. $4\,(z_1 + z_2)/(z_1 - z_2)$

16–20 Let $z = x + iy$. Showing details, find, in terms of x and y:

16. $\text{Im}\,(1/z)$, $\text{Im}\,(1/z^2)$

17. $\text{Re}\,z^4 - (\text{Re}\,z^2)^2$

18. $\text{Re}\,[(1 + i)^{16} z^2]$

19. $\text{Re}\,(z/\bar{z})$, $\text{Im}\,(z/\bar{z})$

20. $\text{Im}\,(1/\bar{z}^2)$

13.2 Polar Form of Complex Numbers. Powers and Roots

We gain further insight into the arithmetic operations of complex numbers if, in addition to the xy-coordinates in the complex plane, we also employ the usual polar coordinates r, θ defined by

(1) $$x = r\cos\theta, \qquad y = r\sin\theta.$$

We see that then $z = x + iy$ takes the so-called **polar form**

(2) $$z = r(\cos\theta + i\sin\theta).$$

r is called the **absolute value** or **modulus** of z and is denoted by $|z|$. Hence

(3) $$|z| = r = \sqrt{x^2 + y^2} = \sqrt{z\bar{z}}.$$

Geometrically, $|z|$ is the distance of the point z from the origin (Fig. 323). Similarly, $|z_1 - z_2|$ is the distance between z_1 and z_2 (Fig. 324).

θ is called the **argument** of z and is denoted by $\arg z$. Thus $\theta = \arg z$ and (Fig. 323)

(4) $$\tan\theta = \frac{y}{x} \qquad (z \neq 0).$$

Geometrically, θ is the directed angle from the positive x-axis to OP in Fig. 323. Here, as in calculus, all ***angles are measured in radians and positive in the counterclockwise sense***.

For $z = 0$ this angle θ is undefined. (Why?) For a given $z \neq 0$ it is determined only up to integer multiples of 2π since cosine and sine are periodic with period 2π. But one often wants to specify a unique value of arg z of a given $z \neq 0$. For this reason one defines the **principal value** Arg z (with capital A!) of arg z by the double inequality

$$-\pi < \text{Arg } z \leqq \pi. \tag{5}$$

Then we have Arg $z = 0$ for positive real $z = x$, which is practical, and Arg $z = \pi$ (not $-\pi$!) for negative real z, e.g., for $z = -4$. The principal value (5) will be important in connection with roots, the complex logarithm (Sec. 13.7), and certain integrals. Obviously, for a given $z \neq 0$, the other values of arg z are arg z = Arg $z \pm 2n\pi$ $(n = \pm 1, \pm 2, \cdots)$.

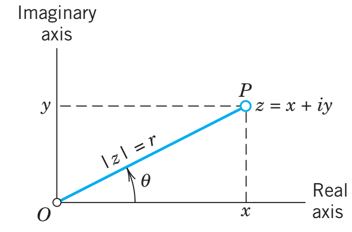


Fig. 323. Complex plane, polar form of a complex number

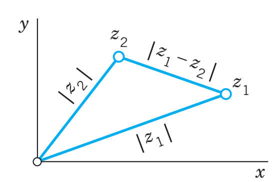


Fig. 324. Distance between two points in the complex plane

EXAMPLE 1 Polar Form of Complex Numbers. Principal Value Arg z

$z = 1 + i$ (Fig. 325) has the polar form $z = \sqrt{2}\,(\cos \frac{1}{4}\pi + i \sin \frac{1}{4}\pi)$. Hence we obtain

$$|z| = \sqrt{2}, \quad \arg z = \tfrac{1}{4}\pi \pm 2n\pi \ (n = 0, 1, \cdots), \qquad \text{and} \qquad \text{Arg } z = \tfrac{1}{4}\pi \quad \text{(the principal value).}$$

Similarly, $z = 3 + 3\sqrt{3}i = 6\,(\cos \frac{1}{3}\pi + i \sin \frac{1}{3}\pi)$, $|z| = 6$, and Arg $z = \frac{1}{3}\pi$. ■

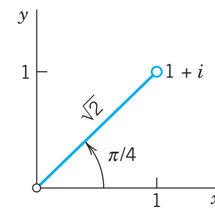


Fig. 325. Example 1

CAUTION! In using (4), we must pay attention to the quadrant in which z lies, since $\tan \theta$ has period π, so that the arguments of z and $-z$ have the same tangent. *Example:* for $\theta_1 = \arg(1 + i)$ and $\theta_2 = \arg(-1 - i)$ we have $\tan \theta_1 = \tan \theta_2 = 1$.

Triangle Inequality

Inequalities such as $x_1 < x_2$ make sense for *real* numbers, but not in complex because *there is no natural way of ordering complex numbers*. However, inequalities between absolute values (which are real!), such as $|z_1| < |z_2|$ (meaning that z_1 is closer to the origin than z_2) are of great importance. The daily bread of the complex analyst is the **triangle inequality**

$$|z_1 + z_2| \leqq |z_1| + |z_2| \tag{6}$$ (Fig. 326)

which we shall use quite frequently. This inequality follows by noting that the three points 0, z_1, and $z_1 + z_2$ are the vertices of a triangle (Fig. 326) with sides $|z_1|$, $|z_2|$, and $|z_1 + z_2|$, and one side cannot exceed the sum of the other two sides. A formal proof is left to the reader (Prob. 33). (The triangle degenerates if z_1 and z_2 lie on the same straight line through the origin.)

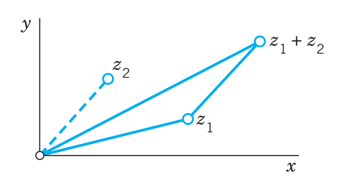


Fig. 326. Triangle inequality

By induction we obtain from (6) the **generalized triangle inequality**

$$|z_1 + z_2 + \cdots + z_n| \leqq |z_1| + |z_2| + \cdots + |z_n|; \qquad (6^*)$$

that is, *the absolute value of a sum cannot exceed the sum of the absolute values of the terms.*

EXAMPLE 2 Triangle Inequality

If $z_1 = 1 + i$ and $z_2 = -2 + 3i$, then (sketch a figure!)

$$|z_1 + z_2| = |-1 + 4i| = \sqrt{17} = 4.123 < \sqrt{2} + \sqrt{13} = 5.020.$$ ■

Multiplication and Division in Polar Form

This will give us a "geometrical" understanding of multiplication and division. Let

$$z_1 = r_1(\cos\theta_1 + i\sin\theta_1) \qquad \text{and} \qquad z_2 = r_2(\cos\theta_2 + i\sin\theta_2).$$

Multiplication. By (3) in Sec. 13.1 the product is at first

$$z_1 z_2 = r_1 r_2[(\cos\theta_1\cos\theta_2 - \sin\theta_1\sin\theta_2) + i(\sin\theta_1\cos\theta_2 + \cos\theta_1\sin\theta_2)].$$

The addition rules for the sine and cosine [(6) in App. A3.1] now yield

$$z_1 z_2 = r_1 r_2[\cos(\theta_1 + \theta_2) + i\sin(\theta_1 + \theta_2)]. \qquad (7)$$

Taking absolute values on both sides of (7), we see that *the absolute value of a product equals the* ***product*** *of the absolute values of the factors*,

$$|z_1 z_2| = |z_1||z_2|. \qquad (8)$$

Taking arguments in (7) shows that *the argument of a product equals the* ***sum*** *of the arguments of the factors*,

$$\arg(z_1 z_2) = \arg z_1 + \arg z_2 \qquad \text{(up to multiples of } 2\pi\text{)}. \qquad (9)$$

Division. We have $z_1 = (z_1/z_2)z_2$. Hence $|z_1| = |(z_1/z_2)z_2| = |z_1/z_2||z_2|$ and by division by $|z_2|$

$$\left|\frac{z_1}{z_2}\right| = \frac{|z_1|}{|z_2|} \qquad (z_2 \neq 0). \qquad (10)$$

Similarly, $\arg z_1 = \arg [(z_1/z_2)z_2] = \arg (z_1/z_2) + \arg z_2$ and by subtraction of $\arg z_2$

(11) $$\arg \frac{z_1}{z_2} = \arg z_1 - \arg z_2 \qquad \text{(up to multiples of } 2\pi).$$

Combining (10) and (11) we also have the analog of (7),

(12) $$\frac{z_1}{z_2} = \frac{r_1}{r_2} [\cos (\theta_1 - \theta_2) + i \sin (\theta_1 - \theta_2)].$$

To comprehend this formula, note that it is the polar form of a complex number of absolute value r_1/r_2 and argument $\theta_1 - \theta_2$. But these are the absolute value and argument of z_1/z_2, as we can see from (10), (11), and the polar forms of z_1 and z_2.

EXAMPLE 3 **Illustration of Formulas (8)–(11)**

Let $z_1 = -2 + 2i$ and $z_2 = 3i$. Then $z_1 z_2 = -6 - 6i$, $z_1/z_2 = \frac{2}{3} + (\frac{2}{3})i$. Hence (make a sketch)

$$|z_1 z_2| = 6\sqrt{2} = 3\sqrt{8} = |z_1||z_2|, \qquad |z_1/z_2| = 2\sqrt{2}/3 = |z_1|/|z_2|,$$

and for the arguments we obtain $\text{Arg } z_1 = 3\pi/4$, $\text{Arg } z_2 = \pi/2$,

$$\text{Arg } (z_1 z_2) = -\frac{3\pi}{4} = \text{Arg } z_1 + \text{Arg } z_2 - 2\pi, \qquad \text{Arg}\left(\frac{z_1}{z_2}\right) = \frac{\pi}{4} = \text{Arg } z_1 - \text{Arg } z_2.$$ ■

EXAMPLE 4 **Integer Powers of *z*. De Moivre's Formula**

From (8) and (9) with $z_1 = z_2 = z$ we obtain by induction for $n = 0, 1, 2, \cdots$

(13) $$z^n = r^n (\cos n\theta + i \sin n\theta).$$

Similarly, (12) with $z_1 = 1$ and $z_2 = z^n$ gives (13) for $n = -1, -2, \cdots$. For $|z| = r = 1$, formula (13) becomes **De Moivre's formula**[3]

(13*) $$(\cos \theta + i \sin \theta)^n = \cos n\theta + i \sin n\theta.$$

We can use this to express $\cos n\theta$ and $\sin n\theta$ in terms of powers of $\cos \theta$ and $\sin \theta$. For instance, for $n = 2$ we have on the left $\cos^2 \theta + 2i \cos \theta \sin \theta - \sin^2 \theta$. Taking the real and imaginary parts on both sides of (13*) with $n = 2$ gives the familiar formulas

$$\cos 2\theta = \cos^2 \theta - \sin^2 \theta, \qquad \sin 2\theta = 2 \cos \theta \sin \theta.$$

This shows that *complex* methods often simplify the derivation of *real* formulas. Try $n = 3$. ■

Roots

If $z = w^n$ $(n = 1, 2, \cdots)$, then to each value of w there corresponds *one* value of z. We shall immediately see that, conversely, to a given $z \neq 0$ there correspond precisely n distinct values of w. Each of these values is called an ***n*th root** of z, and we write

---

[3]ABRAHAM DE MOIVRE (1667–1754), French mathematician, who pioneered the use of complex numbers in trigonometry and also contributed to probability theory (see Sec. 24.8).

$$w = \sqrt[n]{z}. \tag{14}$$

Hence this symbol is ***multivalued***, namely, *n-valued.* The n values of $\sqrt[n]{z}$ can be obtained as follows. We write z and w in polar form

$$z = r(\cos\theta + i\sin\theta) \qquad \text{and} \qquad w = R(\cos\phi + i\sin\phi).$$

Then the equation $w^n = z$ becomes, by De Moivre's formula (with ϕ instead of θ),

$$w^n = R^n(\cos n\phi + i\sin n\phi) = z = r(\cos\theta + i\sin\theta).$$

The absolute values on both sides must be equal; thus, $R^n = r$, so that $R = \sqrt[n]{r}$, where $\sqrt[n]{r}$ is positive real (an absolute value must be nonnegative!) and thus uniquely determined. Equating the arguments $n\phi$ and θ and recalling that θ is determined only up to integer multiples of 2π, we obtain

$$n\phi = \theta + 2k\pi, \qquad \text{thus} \qquad \phi = \frac{\theta}{n} + \frac{2k\pi}{n}$$

where k is an integer. For $k = 0, 1, \cdots, n-1$ we get n *distinct* values of w. Further integers of k would give values already obtained. For instance, $k = n$ gives $2k\pi/n = 2\pi$, hence the w corresponding to $k = 0$, etc. Consequently, $\sqrt[n]{z}$, for $z \neq 0$, has the n distinct values

$$\sqrt[n]{z} = \sqrt[n]{r}\left(\cos\frac{\theta + 2k\pi}{n} + i\sin\frac{\theta + 2k\pi}{n}\right) \tag{15}$$

where $k = 0, 1, \cdots, n-1$. These n values lie on a circle of radius $\sqrt[n]{r}$ with center at the origin and constitute the vertices of a regular polygon of n sides. The value of $\sqrt[n]{z}$ obtained by taking the principal value of arg z and $k = 0$ in (15) is called the **principal value** of $w = \sqrt[n]{z}$.

Taking $z = 1$ in (15), we have $|z| = r = 1$ and Arg $z = 0$. Then (15) gives

$$\sqrt[n]{1} = \cos\frac{2k\pi}{n} + i\sin\frac{2k\pi}{n}, \qquad k = 0, 1, \cdots, n-1. \tag{16}$$

These n values are called the ***n*th roots of unity**. They lie on the circle of radius 1 and center 0, briefly called the **unit circle** (and used quite frequently!). Figures 327–329 show $\sqrt[3]{1} = 1, -\frac{1}{2} \pm \frac{1}{2}\sqrt{3}i$, $\sqrt[4]{1} = \pm 1, \pm i$, and $\sqrt[5]{1}$.

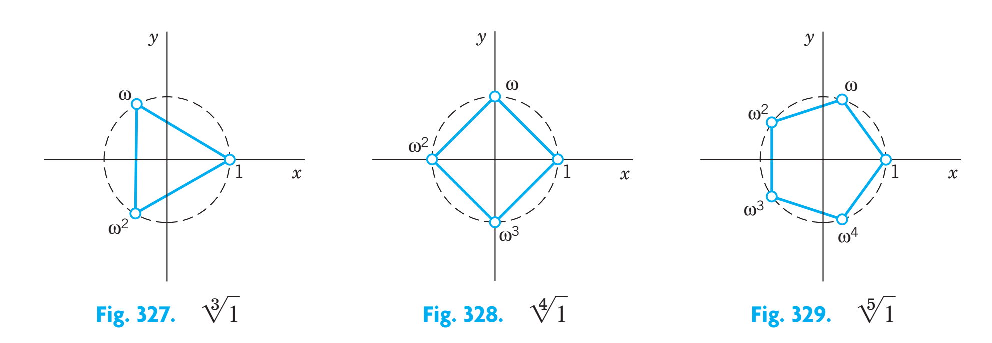


Fig. 327. $\sqrt[3]{1}$ **Fig. 328.** $\sqrt[4]{1}$ **Fig. 329.** $\sqrt[5]{1}$

If ω denotes the value corresponding to $k = 1$ in (16), then the n values of $\sqrt[n]{1}$ can be written as

$$1, \omega, \omega^2, \cdots, \omega^{n-1}.$$

More generally, if w_1 is any nth root of an arbitrary complex number $z\ (\neq 0)$, then the n values of $\sqrt[n]{z}$ in (15) are

$$w_1, \quad w_1\omega, \quad w_1\omega^2, \quad \cdots, \quad w_1\omega^{n-1} \tag{17}$$

because multiplying w_1 by ω^k corresponds to increasing the argument of w_1 by $2k\pi/n$. Formula (17) motivates the introduction of roots of unity and shows their usefulness.

PROBLEM SET 13.2

1–8 POLAR FORM

Represent in polar form and graph in the complex plane as in Fig. 325. Do these problems very carefully because polar forms will be needed frequently. Show the details.

1. $1 + i$

2. $-4 + 4i$

3. $2i, \quad -2i$

4. -5

5. $\dfrac{\sqrt{2} + i/3}{-\sqrt{8} - 2i/3}$

6. $\dfrac{\sqrt{3} - 10i}{-\frac{1}{2}\sqrt{3} + 5i}$

7. $1 + \frac{1}{2}\pi i$

8. $\dfrac{-4 + 19i}{2 + 5i}$

9–14 PRINCIPAL ARGUMENT

Determine the principal value of the argument and graph it as in Fig. 325.

9. $-1 + i$

10. $-5, \quad -5 - i, \quad -5 + i$

11. $3 \pm 4i$

12. $-\pi - \pi i$

13. $(1 + i)^{20}$

14. $-1 + 0.1i, \quad -1 - 0.1i$

15–18 CONVERSION TO $x + iy$

Graph in the complex plane and represent in the form $x + iy$:

15. $3 (\cos \frac{1}{2}\pi - i \sin \frac{1}{2}\pi)$

16. $6 (\cos \frac{1}{3}\pi + i \sin \frac{1}{3}\pi)$

17. $\sqrt{8} (\cos \frac{1}{4}\pi + i \sin \frac{1}{4}\pi)$

18. $\sqrt{50} (\cos \frac{3}{4}\pi + i \sin \frac{3}{4}\pi)$

ROOTS

19. CAS PROJECT. Roots of Unity and Their Graphs. Write a program for calculating these roots and for graphing them as points on the unit circle. Apply the program to $z^n = 1$ with $n = 2, 3, \cdots, 10$. Then extend the program to one for arbitrary roots, using an idea near the end of the text, and apply the program to examples of your choice.

20. TEAM PROJECT. Square Root. (a) Show that $w = \sqrt{z}$ has the values

$$\begin{aligned} w_1 &= \sqrt{r}\left[\cos\frac{\theta}{2} + i \sin\frac{\theta}{2}\right], \\ w_2 &= \sqrt{r}\left[\cos\left(\frac{\theta}{2} + \pi\right) + i \sin\left(\frac{\theta}{2} + \pi\right)\right] \\ &= -w_1. \end{aligned} \tag{18}$$

(b) Obtain from (18) the often more practical formula

$$\sqrt{z} = \pm[\sqrt{\tfrac{1}{2}(|z| + x)} + (\operatorname{sign} y) i \sqrt{\tfrac{1}{2}(|z| + x)}] \tag{19}$$

where sign $y = 1$ if $y \geqq 0$, sign $y = -1$ if $y < 0$, and all square roots of positive numbers are taken with positive sign. *Hint:* Use (10) in App. A3.1 with $x = \theta/2$.

(c) Find the square roots of $-14i$, $-9 - 40i$, and $1 + \sqrt{48}i$ by both (18) and (19) and comment on the work involved.

(d) Do some further examples of your own and apply a method of checking your results.

21–27 ROOTS

Find and graph all roots in the complex plane.

21. $\sqrt[3]{1 + i}$

22. $\sqrt[3]{3 + 4i}$

23. $\sqrt[3]{216}$

24. $\sqrt[4]{-4}$

25. $\sqrt[4]{i}$

26. $\sqrt[8]{1}$

27. $\sqrt[5]{-1}$

28–31 EQUATIONS

Solve and graph the solutions. Show details.

28. $z^2 - (6 - 2i)z + 17 - 6i = 0$

29. $z^2 + z + 1 - i = 0$

30. $z^4 + 324 = 0$. Using the solutions, factor $z^4 + 324$ into quadratic factors with *real* coefficients.

31. $z^4 - 6iz^2 + 16 = 0$

32–35 INEQUALITIES AND EQUALITY

32. Triangle inequality. Verify (6) for $z_1 = 3 + i$, $z_2 = -2 + 4i$

33. Triangle inequality. Prove (6).

34. Re and Im. Prove $|\text{Re } z| \leqq |z|$, $|\text{Im } z| \leqq |z|$.

35. Parallelogram equality. Prove and explain the name

$$|z_1 + z_2|^2 + |z_1 - z_2|^2 = 2\,(|z_1|^2 + |z_2|^2).$$

13.3 Derivative. Analytic Function

Just as the study of calculus or real analysis required concepts such as domain, neighborhood, function, limit, continuity, derivative, etc., so does the study of complex analysis. Since the functions live in the complex plane, the concepts are slightly more difficult or *different* from those in real analysis. This section can be seen as a reference section where many of the concepts needed for the rest of Part D are introduced.

Circles and Disks. Half-Planes

The **unit circle** $|z| = 1$ (Fig. 330) has already occurred in Sec. 13.2. Figure 331 shows a general circle of radius ρ and center a. Its equation is

$$|z - a| = \rho$$

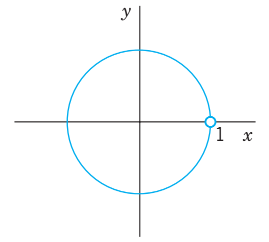


Fig. 330. Unit circle

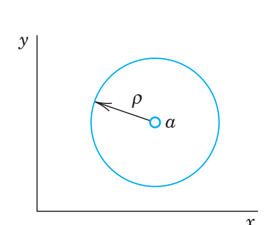


Fig. 331. Circle in the complex plane

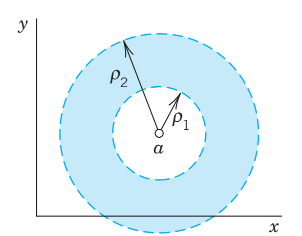


Fig. 332. Annulus in the complex plane

because it is the set of all z whose distance $|z - a|$ from the center a equals ρ. Accordingly, its interior ("**open circular disk**") is given by $|z - a| < \rho$, its interior plus the circle itself ("**closed circular disk**") by $|z - a| \leqq \rho$, and its exterior by $|z - a| > \rho$. As an example, sketch this for $a = 1 + i$ and $\rho = 2$, to make sure that you understand these inequalities.

An open circular disk $|z - a| < \rho$ is also called a **neighborhood** of a or, more precisely, a ρ-*neighborhood* of a. And a has infinitely many of them, one for each value of ρ (> 0), and a is a point of each of them, by definition!

In modern literature *any set* containing a ρ-neighborhood of a is also called a *neighborhood* of a.

Figure 332 shows an **open annulus** (circular ring) $\rho_1 < |z - a| < \rho_2$, which we shall need later. This is the set of all z whose distance $|z - a|$ from a is greater than ρ_1 but less than ρ_2. Similarly, the **closed annulus** $\rho_1 \leqq |z - a| \leqq \rho_2$ includes the two circles.

Half-Planes. By the (open) *upper* **half-plane** we mean the set of all points $z = x + iy$ such that $y > 0$. Similarly, the condition $y < 0$ defines the *lower half-plane*, $x > 0$ the *right half-plane*, and $x < 0$ the *left half-plane*.

For Reference: Concepts on Sets in the Complex Plane

To our discussion of special sets let us add some general concepts related to sets that we shall need throughout Chaps. 13–18; keep in mind that you can find them here.

By a **point set** in the complex plane we mean any sort of collection of finitely many or infinitely many points. Examples are the solutions of a quadratic equation, the points of a line, the points in the interior of a circle as well as the sets discussed just before.

A set S is called **open** if every point of S has a neighborhood consisting entirely of points that belong to S. For example, the points in the interior of a circle or a square form an open set, and so do the points of the right half-plane $\text{Re } z = x > 0$.

A set S is called **connected** if any two of its points can be joined by a chain of finitely many straight-line segments all of whose points belong to S. An open and connected set is called a **domain.** Thus an open disk and an open annulus are domains. An open square with a diagonal removed is not a domain since this set is not connected. (Why?)

The **complement** of a set S in the complex plane is the set of all points of the complex plane that ***do not belong*** to S. A set S is called **closed** if its complement is open. For example, the points on and inside the unit circle form a closed set ("closed unit disk") since its complement $|z| > 1$ is open.

A **boundary point** of a set S is a point every neighborhood of which contains both points that belong to S and points that do not belong to S. For example, the boundary points of an annulus are the points on the two bounding circles. Clearly, if a set S is open, then no boundary point belongs to S; if S is closed, then every boundary point belongs to S. The set of all boundary points of a set S is called the **boundary** of S.

A **region** is a set consisting of a domain plus, perhaps, some or all of its boundary points. WARNING! "Domain" is the *modern* term for an open connected set. Nevertheless, some authors still call a domain a "region" and others make no distinction between the two terms.

Complex Function

Complex analysis is concerned with complex functions that are differentiable in some domain. Hence we should first say what we mean by a complex function and then define the concepts of limit and derivative in complex. This discussion will be similar to that in calculus. Nevertheless it needs great attention because it will show interesting basic differences between real and complex calculus.

Recall from calculus that a *real* function f defined on a set S of real numbers (usually an interval) is a rule that assigns to every x in S a real number $f(x)$, called the *value* of f at x. Now in complex, S is a set of *complex* numbers. And a **function** f defined on S is a rule that assigns to every z in S a complex number w, called the *value* of f at z. We write

$$w = f(z).$$

Here z varies in S and is called a **complex variable**. The set S is called the *domain of definition* of f or, briefly, the **domain** of f. (In most cases S will be open and connected, thus a domain as defined just before.)

Example: $w = f(z) = z^2 + 3z$ is a complex function defined for all z; that is, its domain S is the whole complex plane.

The set of all values of a function f is called the **range** *of* f.

w is complex, and we write $w = u + iv$, where u and v are the real and imaginary parts, respectively. Now w depends on $z = x + iy$. Hence u becomes a real function of x and y, and so does v. We may thus write

$$w = f(z) = u(x, y) + iv(x, y).$$

This shows that a *complex* function $f(z)$ is equivalent to a *pair* of *real* functions $u(x, y)$ and $v(x, y)$, each depending on the two real variables x and y.

EXAMPLE 1 Function of a Complex Variable

Let $w = f(z) = z^2 + 3z$. Find u and v and calculate the value of f at $z = 1 + 3i$.

Solution. $u = \text{Re}\, f(z) = x^2 - y^2 + 3x$ and $v = 2xy + 3y$. Also,

$$f(1 + 3i) = (1 + 3i)^2 + 3(1 + 3i) = 1 - 9 + 6i + 3 + 9i = -5 + 15i.$$

This shows that $u(1, 3) = -5$ and $v(1, 3) = 15$. Check this by using the expressions for u and v. ■

EXAMPLE 2 Function of a Complex Variable

Let $w = f(z) = 2iz + 6\bar{z}$. Find u and v and the value of f at $z = \frac{1}{2} + 4i$.

Solution. $f(z) = 2i(x + iy) + 6(x - iy)$ gives $u(x, y) = 6x - 2y$ and $v(x, y) = 2x - 6y$. Also,

$$f(\tfrac{1}{2} + 4i) = 2i(\tfrac{1}{2} + 4i) + 6(\tfrac{1}{2} - 4i) = i - 8 + 3 - 24i = -5 - 23i.$$

Check this as in Example 1. ■

Remarks on Notation and Terminology

1. Strictly speaking, $f(z)$ denotes the value of f at z, but it is a convenient abuse of language to talk about *the function* $f(z)$ (instead of *the function* f), thereby exhibiting the notation for the independent variable.

2. We assume all functions to be ***single-valued relations***, as usual: to each z in S there corresponds but *one* value $w = f(z)$ (but, of course, several z may give the same value $w = f(z)$, just as in calculus). Accordingly, we shall *not use* the term "multivalued function" (used in some books on complex analysis) for a multivalued relation, in which to a z there corresponds more than one w.

Limit, Continuity

A function $f(z)$ is said to have the **limit** l as z approaches a point z_0, written

$$\lim_{z \to z_0} f(z) = l, \tag{1}$$

if f is defined in a neighborhood of z_0 (except perhaps at z_0 itself) and if the values of f are "close" to l for all z "close" to z_0; in precise terms, if for every positive real ϵ we can find a positive real δ such that for all $z \neq z_0$ in the disk $|z - z_0| < \delta$ (Fig. 333) we have

$$|f(z) - l| < \epsilon; \tag{2}$$

geometrically, if for every $z \neq z_0$ in that δ-disk the value of f lies in the disk (2).

Formally, this definition is similar to that in calculus, but there is a big difference. Whereas in the real case, x can approach an x_0 only along the real line, here, by definition,

z may approach z_0 ***from any direction*** in the complex plane. This will be quite essential in what follows.

If a limit exists, it is unique. (See Team Project 24.)

A function $f(z)$ is said to be **continuous** at $z = z_0$ if $f(z_0)$ is defined and

(3) $$\lim_{z \to z_0} f(z) = f(z_0).$$

Note that by definition of a limit this implies that $f(z)$ is defined in some neighborhood of z_0.

$f(z)$ is said to be *continuous in a domain* if it is continuous at each point of this domain.

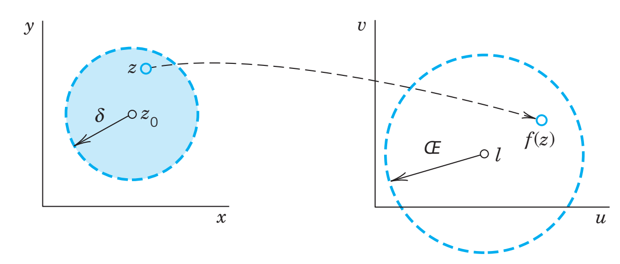


Fig. 333. Limit

Derivative

The **derivative** of a complex function f at a point z_0 is written $f'(z_0)$ and is defined by

(4) $$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}$$

provided this limit exists. Then f is said to be **differentiable** at z_0. If we write $\Delta z = z - z_0$, we have $z = z_0 + \Delta z$ and (4) takes the form

(4′) $$f'(z_0) = \lim_{z \to z_0} \frac{f(z) - f(z_0)}{z - z_0}.$$

Now comes an ***important point***. Remember that, by the definition of limit, $f(z)$ is defined in a neighborhood of z_0 and z in (4′) may approach z_0 from any direction in the complex plane. Hence differentiability at z_0 means that, along whatever path z approaches z_0, the quotient in (4′) always approaches a certain value and all these values are equal. This is important and should be kept in mind.

EXAMPLE 3 **Differentiability. Derivative**

The function $f(z) = z^2$ is differentiable for all z and has the derivative $f'(z) = 2z$ because

$$f'(z) = \lim_{\Delta z \to 0} \frac{(z + \Delta z)^2 - z^2}{\Delta z} = \lim_{\Delta z \to 0} \frac{z^2 + 2z\,\Delta z + (\Delta z)^2 - z^2}{\Delta z} = \lim_{\Delta z \to 0} (2z + \Delta z) = 2z.$$ ■

The ***differentiation rules*** *are the same as in real calculus,* since their proofs are literally the same. Thus for any differentiable functions f and g and constant c we have

$$(cf)' = cf', \quad (f+g)' = f' + g', \quad (fg)' = f'g + fg', \quad \left(\frac{f}{g}\right)' = \frac{f'g - fg'}{g^2}$$

as well as the chain rule and the power rule $(z^n)' = nz^{n-1}$ (n integer).

Also, if $f(z)$ is differentiable at z_0, it is continuous at z_0. (See Team Project 24.)

EXAMPLE 4 $\bar{z}$ **not Differentiable**

It may come as a surprise that there are many complex functions that do not have a derivative at any point. For instance, $f(z) = \bar{z} = x - iy$ is such a function. To see this, we write $\Delta z = \Delta x + i\Delta y$ and obtain

$$\frac{f(z+\Delta z) - f(z)}{\Delta z} = \frac{\overline{(z+\Delta z)} - \bar{z}}{\Delta z} = \frac{\overline{\Delta z}}{\Delta z} = \frac{\Delta x - i\Delta y}{\Delta x + i\Delta y}. \tag{5}$$

If $\Delta y = 0$, this is $+1$. If $\Delta x = 0$, this is -1. Thus (5) approaches $+1$ along path I in Fig. 334 but -1 along path II. Hence, by definition, the limit of (5) as $\Delta z \to 0$ does not exist at any z. ■

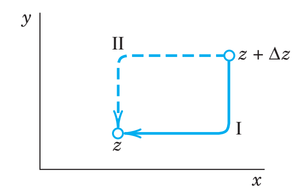


Fig. 334. Paths in (5)

Surprising as Example 4 may be, it merely illustrates that differentiability of a *complex* function is a rather severe requirement.

The idea of proof (approach of z from different directions) is basic and will be used again as the crucial argument in the next section.

Analytic Functions

Complex analysis is concerned with the theory and application of "analytic functions," that is, functions that are differentiable in some domain, so that we can do "calculus in complex." The definition is as follows.

DEFINITION

Analyticity

A function $f(z)$ is said to be *analytic in a domain D* if $f(z)$ is defined and differentiable at all points of D. The function $f(z)$ is said to be *analytic at a point* $z = z_0$ in D if $f(z)$ is analytic in a neighborhood of z_0.

Also, by an **analytic function** we mean a function that is analytic in *some* domain.

Hence analyticity of $f(z)$ at z_0 means that $f(z)$ has a derivative at every point in some neighborhood of z_0 (including z_0 itself since, by definition, z_0 is a point of all its neighborhoods). This concept is *motivated* by the fact that it is of no practical interest if a function is differentiable merely at a single point z_0 but not throughout some neighborhood of z_0. Team Project 24 gives an example.

A more modern term for *analytic in D* is *holomorphic in D*.

EXAMPLE 5 Polynomials, Rational Functions

The nonnegative integer powers $1, z, z^2, \cdots$ are analytic in the entire complex plane, and so are **polynomials**, that is, functions of the form

$$f(z) = c_0 + c_1 z + c_2 z^2 + \cdots + c_n z^n$$

where $c_0, \cdots, c_n$ are complex constants.

The quotient of two polynomials $g(z)$ and $h(z)$,

$$f(z) = \frac{g(z)}{h(z)},$$

is called a **rational function**. This f is analytic except at the points where $h(z) = 0$; here we assume that common factors of g and h have been canceled.

Many further analytic functions will be considered in the next sections and chapters. ■

The concepts discussed in this section extend familiar concepts of calculus. Most important is the concept of an analytic function, the exclusive concern of complex analysis. Although many simple functions are not analytic, the large variety of remaining functions will yield a most beautiful branch of mathematics that is very useful in engineering and physics.

PROBLEM SET 13.3

1–8 REGIONS OF PRACTICAL INTEREST

Determine and sketch or graph the sets in the complex plane given by

1. $|z + 1 - 5i| \leqq \frac{3}{2}$
2. $0 < |z| < 1$
3. $\pi < |z - 4 + 2i| < 3\pi$
4. $-\pi < \text{Im } z < \pi$
5. $|\arg z| < \frac{1}{4}\pi$
6. $\text{Re}(1/z) < 1$
7. $\text{Re } z \geqq -1$
8. $|z + i| \geqq |z - i|$
9. **WRITING PROJECT. Sets in the Complex Plane.** Write a report by formulating the corresponding portions of the text in your own words and illustrating them with examples of your own.

COMPLEX FUNCTIONS AND THEIR DERIVATIVES

10–12 Function Values. Find Re f, and Im f and their values at the given point z.

10. $f(z) = 5z^2 - 12z + 3 + 2i$ at $4 - 3i$
11. $f(z) = 1/(1 - z)$ at $1 - i$
12. $f(z) = (z - 2)/(z + 2)$ at $8i$
13. **CAS PROJECT. Graphing Functions.** Find and graph Re f, Im f, and $|f|$ as surfaces over the z-plane. Also graph the two families of curves Re $f(z)$ = const and Im $f(z)$ = const in the same figure, and the curves $|f(z)|$ = const in another figure, where **(a)** $f(z) = z^2$, **(b)** $f(z) = 1/z$, (c) $f(z) = z^4$.

14–17 Continuity. Find out, and give reason, whether $f(z)$ is continuous at $z = 0$ if $f(0) = 0$ and for $z \neq 0$ the function f is equal to:

14. $(\text{Re } z^2)/|z|$
15. $|z|^2 \text{ Im}(1/z)$
16. $(\text{Im } z^2)/|z|^2$
17. $(\text{Re } z)/(1 - |z|)$

18–23 Differentiation. Find the value of the derivative of

18. $(z - i)/(z + i)$ at i
19. $(z - 4i)^8$ at $= 3 + 4i$
20. $(1.5z + 2i)/(3iz - 4)$ at any z. Explain the result.
21. $i(1 - z)^n$ at 0
22. $(iz^3 + 3z^2)^3$ at $2i$
23. $z^3/(z + i)^3$ at i
24. **TEAM PROJECT. Limit, Continuity, Derivative**

    (a) Limit. Prove that (1) is equivalent to the pair of relations

    $$\lim_{z \to z_0} \text{Re } f(z) = \text{Re } l, \qquad \lim_{z \to z_0} \text{Im } f(z) = \text{Im } l.$$

    (b) Limit. If $\lim_{z \to z_0} f(x)$ exists, show that this limit is unique.

    (c) Continuity. If $z_1, z_2, \cdots$ are complex numbers for which $\lim_{n \to \infty} z_n = a$, and if $f(z)$ is continuous at $z = a$, show that $\lim_{n \to \infty} f(z_n) = f(a)$.

(d) Continuity. If $f(z)$ is differentiable at z_0, show that $f(z)$ is continuous at z_0.

(e) Differentiability. Show that $f(z) = \text{Re } z = x$ is not differentiable at any z. Can you find other such functions?

(f) Differentiability. Show that $f(z) = |z|^2$ is differentiable only at $z = 0$; hence it is nowhere analytic.

25. WRITING PROJECT. Comparison with Calculus. Summarize the second part of this section beginning with *Complex Function*, and indicate what is conceptually analogous to calculus and what is not.

13.4 Cauchy–Riemann Equations. Laplace's Equation

As we saw in the last section, to do complex analysis (i.e., "calculus in the complex") on any complex function, we require that function to be *analytic on some domain* that is differentiable in that domain.

The Cauchy–Riemann equations are the most important equations in this chapter and one of the pillars on which complex analysis rests. They provide a criterion (a test) for the analyticity of a complex function

$$w = f(z) = u(x, y) + iv(x, y).$$

Roughly, f is analytic in a domain D if and only if the first partial derivatives of u and v satisfy the two **Cauchy–Riemann equations**[4]

$$u_x = v_y, \qquad u_y = -v_x \tag{1}$$

everywhere in D; here $u_x = \partial u/\partial x$ and $u_y = \partial u/\partial y$ (and similarly for v) are the usual notations for partial derivatives. The precise formulation of this statement is given in Theorems 1 and 2.

Example: $f(z) = z^2 = x^2 - y^2 + 2ixy$ is analytic for all z (see Example 3 in Sec. 13.3), and $u = x^2 - y^2$ and $v = 2xy$ satisfy (1), namely, $u_x = 2x = v_y$ as well as $u_y = -2y = -v_x$. More examples will follow.

THEOREM 1

Cauchy–Riemann Equations

Let $f(z) = u(x, y) + iv(x, y)$ be defined and continuous in some neighborhood of a point $z = x + iy$ and differentiable at z itself. Then, at that point, the first-order partial derivatives of u and v exist and satisfy the Cauchy–Riemann equations (1).

Hence, if $f(z)$ is analytic in a domain D, those partial derivatives exist and satisfy (1) *at all points of D.*

---

[4] The French mathematician AUGUSTIN-LOUIS CAUCHY (see Sec. 2.5) and the German mathematicians BERNHARD RIEMANN (1826–1866) and KARL WEIERSTRASS (1815–1897; see also Sec. 15.5) are the founders of complex analysis. Riemann received his Ph.D. (in 1851) under Gauss (Sec. 5.4) at Göttingen, where he also taught until he died, when he was only 39 years old. He introduced the concept of the integral as it is used in basic calculus courses, and made important contributions to differential equations, number theory, and mathematical physics. He also developed the so-called Riemannian geometry, which is the mathematical foundation of Einstein's theory of relativity; see Ref. [GenRef9] in App. 1.

PROOF By assumption, the derivative $f'(z)$ at z exists. It is given by

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z}. \tag{2}$$

The idea of the proof is very simple. By the definition of a limit in complex (Sec. 13.3), we can let Δz approach zero along any path in a neighborhood of z. Thus we may choose the two paths I and II in Fig. 335 and equate the results. By comparing the real parts we shall obtain the first Cauchy–Riemann equation and by comparing the imaginary parts the second. The technical details are as follows.

We write $\Delta z = \Delta x + i\,\Delta y$. Then $z + \Delta z = x + \Delta x + i(y + \Delta y)$, and in terms of u and v the derivative in (2) becomes

$$f'(z) = \lim_{\Delta z \to 0} \frac{[u(x + \Delta x, y + \Delta y) + iv(x + \Delta x, y + \Delta y)] - [u(x, y) + iv(x, y)]}{\Delta x + i\,\Delta y}. \tag{3}$$

We first choose path I in Fig. 335. Thus we let $\Delta y \to 0$ first and then $\Delta x \to 0$. After Δy is zero, $\Delta z = \Delta x$. Then (3) becomes, if we first write the two u-terms and then the two v-terms,

$$f'(z) = \lim_{\Delta x \to 0} \frac{u(x + \Delta x, y) - u(x, y)}{\Delta x} + i \lim_{\Delta x \to 0} \frac{v(x + \Delta x, y) - v(x, y)}{\Delta x}.$$

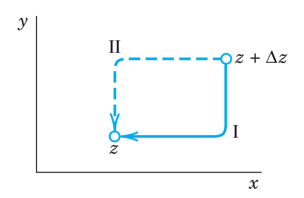


Fig. 335. Paths in (2)

Since $f'(z)$ exists, the two real limits on the right exist. By definition, they are the partial derivatives of u and v with respect to x. Hence the derivative $f'(z)$ of $f(z)$ can be written

$$f'(z) = u_x + iv_x. \tag{4}$$

Similarly, if we choose path II in Fig. 335, we let $\Delta x \to 0$ first and then $\Delta y \to 0$. After Δx is zero, $\Delta z = i\,\Delta y$, so that from (3) we now obtain

$$f'(z) = \lim_{\Delta y \to 0} \frac{u(x, y + \Delta y) - u(x, y)}{i\,\Delta y} + i \lim_{\Delta y \to 0} \frac{v(x, y + \Delta y) - v(x, y)}{i\,\Delta y}.$$

Since $f'(z)$ exists, the limits on the right exist and give the partial derivatives of u and v with respect to y; noting that $1/i = -i$, we thus obtain

$$f'(z) = -iu_y + v_y. \tag{5}$$

The existence of the derivative $f'(z)$ thus implies the existence of the four partial derivatives in (4) and (5). By equating the real parts u_x and v_y in (4) and (5) we obtain the first

Cauchy–Riemann equation (1). Equating the imaginary parts gives the other. This proves the first statement of the theorem and implies the second because of the definition of analyticity. ■

Formulas (4) and (5) are also quite practical for calculating derivatives $f'(z)$, as we shall see.

EXAMPLE 1 **Cauchy–Riemann Equations**

$f(z) = z^2$ is analytic for all z. It follows that the Cauchy–Riemann equations must be satisfied (as we have verified above).

For $f(z) = \bar{z} = x - iy$ we have $u = x, v = -y$ and see that the second Cauchy–Riemann equation is satisfied, $u_y = -v_x = 0$, but the first is not: $u_x = 1 \neq v_y = -1$. We conclude that $f(z) = \bar{z}$ is not analytic, confirming Example 4 of Sec. 13.3. Note the savings in calculation! ■

The Cauchy–Riemann equations are fundamental because they are not only necessary but also sufficient for a function to be analytic. More precisely, the following theorem holds.

THEOREM 2 **Cauchy–Riemann Equations**

If two real-valued continuous functions $u(x, y)$ and $v(x, y)$ of two real variables x and y have ***continuous*** *first partial derivatives that satisfy the Cauchy–Riemann equations in some domain D, then the complex function $f(z) = u(x, y) + iv(x, y)$ is analytic in D.*

The proof is more involved than that of Theorem 1 and we leave it optional (see App. 4).

Theorems 1 and 2 are of great practical importance, since, by using the Cauchy–Riemann equations, we can now easily find out whether or not a given complex function is analytic.

EXAMPLE 2 **Cauchy–Riemann Equations. Exponential Function**

Is $f(z) = u(x, y) + iv(x, y) = e^x(\cos y + i \sin y)$ analytic?

Solution. We have $u = e^x \cos y, v = e^x \sin y$ and by differentiation

$$u_x = e^x \cos y, \qquad v_y = e^x \cos y$$
$$u_y = -e^x \sin y, \qquad v_x = e^x \sin y.$$

We see that the Cauchy–Riemann equations are satisfied and conclude that $f(z)$ is analytic for all z. ($f(z)$ will be the complex analog of e^x known from calculus.) ■

EXAMPLE 3 **An Analytic Function of Constant Absolute Value Is Constant**

The Cauchy–Riemann equations also help in deriving general properties of analytic functions.

For instance, show that if $f(z)$ is analytic in a domain D and $|f(z)| = k = \text{const}$ in D, then $f(z) = \text{const}$ in D. (We shall make crucial use of this in Sec. 18.6 in the proof of Theorem 3.)

Solution. By assumption, $|f|^2 = |u + iv|^2 = u^2 + v^2 = k^2$. By differentiation,

$$uu_x + vv_x = 0,$$
$$uu_y + vv_y = 0.$$

Now use $v_x = -u_y$ in the first equation and $v_y = u_x$ in the second, to get

$$\text{(6)} \qquad \begin{aligned} &\text{(a)} \quad uu_x - vu_y = 0, \\ &\text{(b)} \quad uu_y - vu_x = 0. \end{aligned}$$

To get rid of u_y, multiply (6a) by u and (6b) by v and add. Similarly, to eliminate u_x, multiply (6a) by $-v$ and (6b) by u and add. This yields

$$(u^2 + v^2)u_x = 0,$$
$$(u^2 + v^2)u_y = 0.$$

If $k^2 = u^2 + v^2 = 0$, then $u = v = 0$; hence $f = 0$. If $k^2 = u^2 + v^2 \neq 0$, then $u_x = u_y = 0$. Hence, by the Cauchy–Riemann equations, also $u_x = v_y = 0$. Together this implies $u = \text{const}$ and $v = \text{const}$; hence $f = \text{const}$. ■

We mention that, if we use the polar form $z = r(\cos\theta + i\sin\theta)$ and set $f(z) = u(r, \theta) + iv(r, \theta)$, then the **Cauchy–Riemann equations** are (Prob. 1)

(7)
$$u_r = \frac{1}{r} v_\theta, \qquad v_r = -\frac{1}{r} u_\theta \qquad (r > 0).$$

Laplace's Equation. Harmonic Functions

The great importance of complex analysis in engineering mathematics results mainly from the fact that both the real part and the imaginary part of an analytic function satisfy Laplace's equation, the most important PDE of physics. It occurs in gravitation, electrostatics, fluid flow, heat conduction, and other applications (see Chaps. 12 and 18).

THEOREM 3

Laplace's Equation

If $f(z) = u(x, y) + iv(x, y)$ is analytic in a domain D, then both u and v satisfy **Laplace's equation**

(8) $$\nabla^2 u = u_{xx} + u_{yy} = 0$$

(∇^2 read "nabla squared") *and*

(9) $$\nabla^2 v = v_{xx} + v_{yy} = 0,$$

in D and have continuous second partial derivatives in D.

PROOF Differentiating $u_x = v_y$ with respect to x and $u_y = -v_x$ with respect to y, we have

(10) $$u_{xx} = v_{yx}, \qquad u_{yy} = -v_{xy}.$$

Now the derivative of an analytic function is itself analytic, as we shall prove later (in Sec. 14.4). This implies that u and v have continuous partial derivatives of all orders; in particular, the mixed second derivatives are equal: $v_{yx} = v_{xy}$. By adding (10) we thus obtain (8). Similarly, (9) is obtained by differentiating $u_x = v_y$ with respect to y and $u_y = -v_x$ with respect to x and subtracting, using $u_{xy} = u_{yx}$. ■

Solutions of Laplace's equation having ***continuous*** second-order partial derivatives are called **harmonic functions** and their theory is called **potential theory** (see also Sec. 12.11). Hence the real and imaginary parts of an analytic function are harmonic functions.

If two harmonic functions u and v satisfy the Cauchy–Riemann equations in a domain D, they are the real and imaginary parts of an analytic function f in D. Then v is said to be a **harmonic conjugate function** of u in D. (Of course, this has absolutely nothing to do with the use of "conjugate" for $\bar{z}$.)

EXAMPLE 4 How to Find a Harmonic Conjugate Function by the Cauchy–Riemann Equations

Verify that $u = x^2 - y^2 - y$ is harmonic in the whole complex plane and find a harmonic conjugate function v of u.

Solution. $\nabla^2 u = 0$ by direct calculation. Now $u_x = 2x$ and $u_y = -2y - 1$. Hence because of the Cauchy–Riemann equations a conjugate v of u must satisfy

$$v_y = u_x = 2x, \qquad v_x = -u_y = 2y + 1.$$

Integrating the first equation with respect to y and differentiating the result with respect to x, we obtain

$$v = 2xy + h(x), \qquad v_x = 2y + \frac{dh}{dx}.$$

A comparison with the second equation shows that $dh/dx = 1$. This gives $h(x) = x + c$. Hence $v = 2xy + x + c$ (c any real constant) is the most general harmonic conjugate of the given u. The corresponding analytic function is

$$f(z) = u + iv = x^2 - y^2 - y + i(2xy + x + c) = z^2 + iz + ic. \qquad \blacksquare$$

Example 4 illustrates that *a conjugate of a given harmonic function is uniquely determined up to an arbitrary real additive constant.*

The Cauchy–Riemann equations are the most important equations in this chapter. Their relation to Laplace's equation opens a wide range of engineering and physical applications, as shown in Chap. 18.

PROBLEM SET 13.4

1. **Cauchy–Riemann equations in polar form.** Derive (7) from (1).

2–11 CAUCHY–RIEMANN EQUATIONS

Are the following functions analytic? Use (1) or (7).

2. $f(z) = iz\bar{z}$
3. $f(z) = e^{-2x}(\cos 2y - i \sin 2y)$
4. $f(z) = e^x(\cos y - i \sin y)$
5. $f(z) = \text{Re}(z^2) - i\,\text{Im}(z^2)$
6. $f(z) = 1/(z - z^5)$
7. $f(z) = i/z^8$
8. $f(z) = \text{Arg}\, 2\pi z$
9. $f(z) = 3\pi^2/(z^3 + 4\pi^2 z)$
10. $f(z) = \ln |z| + i\,\text{Arg}\, z$
11. $f(z) = \cos x \cosh y - i \sin x \sinh y$

12–19 HARMONIC FUNCTIONS

Are the following functions harmonic? If your answer is yes, find a corresponding analytic function $f(z) = u(x, y) + iv(x, y)$.

12. $u = x^2 + y^2$
13. $u = xy$
14. $v = xy$
15. $u = x/(x^2 + y^2)$
16. $u = \sin x \cosh y$
17. $v = (2x + 1)y$
18. $u = x^3 - 3xy^2$
19. $v = e^x \sin 2y$
20. **Laplace's equation.** Give the details of the derivative of (9).

21–24

Determine a and b so that the given function is harmonic and find a harmonic conjugate.

21. $u = e^{\pi x} \cos av$
22. $u = \cos ax \cosh 2y$
23. $u = ax^3 + bxy$
24. $u = \cosh ax \cos y$
25. **CAS PROJECT. Equipotential Lines.** Write a program for graphing equipotential lines u = const of a harmonic function u and of its conjugate v on the same axes. Apply the program to **(a)** $u = x^2 - y^2$, $v = 2xy$, **(b)** $u = x^3 - 3xy^2$, $v = 3x^2y - y^3$.
26. Apply the program in Prob. 25 to $u = e^x \cos y$, $v = e^x \sin y$ and to an example of your own.

27. Harmonic conjugate. Show that if u is harmonic and v is a harmonic conjugate of u, then u is a harmonic conjugate of $-v$.

28. Illustrate Prob. 27 by an example.

29. Two further formulas for the derivative. Formulas (4), (5), and (11) (below) are needed from time to time. Derive

$$f'(z) = u_x - iu_y, \qquad f'(z) = v_y + iv_x. \tag{11}$$

30. TEAM PROJECT. Conditions for $f(z)$ = const. Let $f(z)$ be analytic. Prove that each of the following conditions is sufficient for $f(z)$ = const.

(a) $\text{Re} f(z) = \text{const}$

(b) $\text{Im} f(z) = \text{const}$

(c) $f'(z) = 0$

(d) $|f(z)| = \text{const}$ (see Example 3)

13.5 Exponential Function

In the remaining sections of this chapter we discuss the basic elementary complex functions, the exponential function, trigonometric functions, logarithm, and so on. They will be counterparts to the familiar functions of calculus, to which they reduce when $z = x$ is real. They are indispensable throughout applications, and some of them have interesting properties not shared by their real counterparts.

We begin with one of the most important analytic functions, the complex **exponential function**

$$e^z, \qquad \text{also written} \qquad \exp z.$$

The definition of e^z in terms of the real functions e^x, $\cos y$, and $\sin y$ is

$$e^z = e^x(\cos y + i \sin y). \tag{1}$$

This definition is motivated by the fact the e^z ***extends*** the real exponential function e^x of calculus in a natural fashion. Namely:

(A) $e^z = e^x$ for real $z = x$ because $\cos y = 1$ and $\sin y = 0$ when $y = 0$.

(B) e^z is analytic for all z. (Proved in Example 2 of Sec. 13.4.)

(C) The derivative of e^z is e^z, that is,

$$(e^z)' = e^z. \tag{2}$$

This follows from (4) in Sec. 13.4,

$$(e^z)' = (e^x \cos y)_x + i(e^x \sin y)_x = e^x \cos y + ie^x \sin y = e^z.$$

REMARK. This definition provides for a relatively simple discussion. We could define e^z by the familiar series $1 + x + x^2/2! + x^3/3! + \cdots$ with x replaced by z, but we would then have to discuss complex series at this very early stage. (We will show the connection in Sec. 15.4.)

Further Properties. A function $f(z)$ that is analytic for all z is called an **entire function**. Thus, e^z is entire. Just as in calculus the ***functional relation***

$$e^{z_1+z_2} = e^{z_1}e^{z_2} \tag{3}$$

holds for any $z_1 = x_1 + iy_1$ and $z_2 = x_2 + iy_2$. Indeed, by (1),

$$e^{z_1}e^{z_2} = e^{x_1}(\cos y_1 + i \sin y_1)e^{x_2}(\cos y_2 + i \sin y_2).$$

Since $e^{x_1}e^{x_2} = e^{x_1+x_2}$ for these *real* functions, by an application of the addition formulas for the cosine and sine functions (similar to that in Sec. 13.2) we see that

$$e^{z_1}e^{z_2} = e^{x_1+x_2}[\cos (y_1 + y_2) + i \sin (y_1 + y_2)] = e^{z_1+z_2}$$

as asserted. An interesting special case of (3) is $z_1 = x$, $z_2 = iy$; then

$$e^z = e^x e^{iy}. \tag{4}$$

Furthermore, for $z = iy$ we have from (1) the so-called **Euler formula**

$$e^{iy} = \cos y + i \sin y. \tag{5}$$

Hence the **polar form** of a complex number, $z = r(\cos \theta + i \sin \theta)$, may now be written

$$z = re^{i\theta}. \tag{6}$$

From (5) we obtain

$$e^{2\pi i} = 1 \tag{7}$$

as well as the important formulas (verify!)

$$e^{\pi i/2} = i, \qquad e^{\pi i} = -1, \qquad e^{-\pi i/2} = -i, \qquad e^{-\pi i} = -1. \tag{8}$$

Another consequence of (5) is

$$|e^{iy}| = |\cos y + i \sin y| = \sqrt{\cos^2 y + \sin^2 y} = 1. \tag{9}$$

That is, for pure imaginary exponents, the exponential function has absolute value 1, a result you should remember. From (9) and (1),

$$|e^z| = e^x. \qquad \text{Hence} \qquad \arg e^z = y \pm 2n\pi \quad (n = 0, 1, 2, \cdots), \tag{10}$$

since $|e^z| = e^x$ shows that (1) is actually e^z in polar form.

From $|e^z| = e^x \neq 0$ in (10) we see that

$$e^x \neq 0 \qquad \text{for all } z. \tag{11}$$

So here we have an entire function that never vanishes, in contrast to (nonconstant) polynomials, which are also entire (Example 5 in Sec. 13.3) but always have a zero, as is proved in algebra.

Periodicity of e^z with period $2\pi i$,

$$e^{z+2\pi i} = e^z \qquad \text{for all } z \tag{12}$$

is a basic property that follows from (1) and the periodicity of cos y and sin y. Hence all the values that $w = e^z$ can assume are already assumed in the horizontal strip of width 2π

$$-\pi < y \leqq \pi \qquad \text{(Fig. 336).} \tag{13}$$

This infinite strip is called a **fundamental region** of e^z.

EXAMPLE 1 Function Values. Solution of Equations

Computation of values from (1) provides no problem. For instance,

$$e^{1.4-0.6i} = e^{1.4}(\cos 0.6 - i \sin 0.6) = 4.055(0.8253 - 0.5646i) = 3.347 - 2.289i$$

$$|e^{1.4-1.6i}| = e^{1.4} = 4.055, \qquad \text{Arg } e^{1.4-0.6i} = -0.6.$$

To illustrate (3), take the product of

$$e^{2+i} = e^2(\cos 1 + i \sin 1) \qquad \text{and} \qquad e^{4-i} = e^4(\cos 1 - i \sin 1)$$

and verify that it equals $e^2e^4(\cos^2 1 + \sin^2 1) = e^6 = e^{(2+i)+(4-i)}$.

To solve the equation $e^z = 3 + 4i$, note first that $|e^z| = e^x = 5, x = \ln 5 = 1.609$ is the real part of all solutions. Now, since $e^x = 5$,

$$e^x \cos y = 3, \qquad e^x \sin y = 4, \qquad \cos y = 0.6, \qquad \sin y = 0.8, \qquad y = 0.927.$$

Ans. $z = 1.609 + 0.927i \pm 2n\pi i \ (n = 0, 1, 2, \cdots)$. These are infinitely many solutions (due to the periodicity of e^z). They lie on the vertical line $x = 1.609$ at a distance 2π from their neighbors. ■

To summarize: many properties of $e^z = \exp z$ parallel those of e^x; an exception is the periodicity of e^z with $2\pi i$, which suggested the concept of a fundamental region. Keep in mind that e^z is an *entire function.* (Do you still remember what that means?)

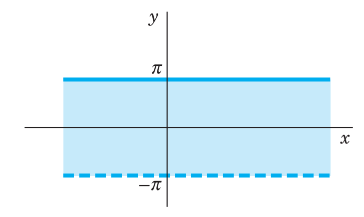


Fig. 336. Fundamental region of the exponential function e^z in the z-plane

PROBLEM SET 13.5

1. e^z is entire. Prove this.

2–7 Function Values. Find e^z in the form $u + iv$ and $|e^z|$ if z equals

2. $3 + 4i$

3. $2\pi i(1 + i)$

4. $0.6 - 1.8i$

5. $2 + 3\pi i$

6. $11\pi i/2$

7. $\sqrt{2} + \frac{1}{2}\pi i$

8–13 Polar Form. Write in exponential form (6):

8. $\sqrt[n]{z}$

9. $4 + 3i$

10. $\sqrt{i}, \quad \sqrt{-i}$

11. -6.3

12. $1/(1 - z)$

13. $1 + i$

14–17 Real and Imaginary Parts. Find Re and Im of

14. $e^{-\pi z}$

15. $\exp(z^2)$

16. $e^{1/z}$

17. $\exp(z^3)$

18. TEAM PROJECT. Further Properties of the Exponential Function. (a) Analyticity. Show that e^z is entire. What about $e^{1/z}$? $e^{\bar{z}}$? $e^x(\cos ky + i \sin ky)$? (Use the Cauchy–Riemann equations.)

(b) Special values. Find all z such that (i) e^z is real, (ii) $|e^{-\bar{z}}| < 1$, (iii) $e^{\bar{z}} = \overline{e^z}$.

(c) Harmonic function. Show that $u = e^{xy} \cos(x^2/2 - y^2/2)$ is harmonic and find a conjugate.

(d) Uniqueness. It is interesting that $f(z) = e^z$ is uniquely determined by the two properties $f(x + i0) = e^x$ and $f'(z) = f(z)$, where f is assumed to be entire. Prove this using the Cauchy–Riemann equations.

19–22 **Equations.** Find all solutions and graph some of them in the complex plane.

19. $e^z = 1$

20. $e^z = 4 + 3i$

21. $e^z = 0$

22. $e^z = -2$

13.6 Trigonometric and Hyperbolic Functions. Euler's Formula

Just as we extended the real e^x to the complex e^z in Sec. 13.5, we now want to extend the familiar *real* trigonometric functions to *complex trigonometric functions.* We can do this by the use of the Euler formulas (Sec. 13.5)

$$e^{ix} = \cos x + i \sin x, \qquad e^{-ix} = \cos x - i \sin x.$$

By addition and subtraction we obtain for the *real* cosine and sine

$$\cos x = \tfrac{1}{2}(e^{ix} + e^{-ix}), \qquad \sin x = \frac{1}{2i}(e^{ix} - e^{-ix}).$$

This suggests the following definitions for complex values $z = x + iy$:

$$\cos z = \tfrac{1}{2}(e^{iz} + e^{-iz}), \qquad \sin z = \frac{1}{2i}(e^{iz} - e^{-iz}). \tag{1}$$

It is quite remarkable that here in complex, functions come together that are unrelated in real. This is not an isolated incident but is typical of the general situation and shows the advantage of working in complex.

Furthermore, as in calculus we define

$$\tan z = \frac{\sin z}{\cos z}, \qquad \cot z = \frac{\cos z}{\sin z} \tag{2}$$

and

$$\sec z = \frac{1}{\cos z}, \qquad \csc z = \frac{1}{\sin z}. \tag{3}$$

Since e^z is entire, $\cos z$ and $\sin z$ are entire functions. $\tan z$ and $\sec z$ are not entire; they are analytic except at the points where $\cos z$ is zero; and $\cot z$ and $\csc z$ are analytic except

where $\sin z$ is zero. Formulas for the derivatives follow readily from $(e^z)' = e^z$ and (1)–(3); as in calculus,

$$(4) \qquad (\cos z)' = -\sin z, \qquad (\sin z)' = \cos z, \qquad (\tan z)' = \sec^2 z,$$

etc. Equation (1) also shows that **Euler's formula *is valid in complex***:

$$\textbf{(5)} \qquad e^{iz} = \cos z + i \sin z \qquad \text{for all } z.$$

The real and imaginary parts of $\cos z$ and $\sin z$ are needed in computing values, and they also help in displaying properties of our functions. We illustrate this with a typical example.

EXAMPLE 1 Real and Imaginary Parts. Absolute Value. Periodicity

Show that

$$\textbf{(6)} \qquad \begin{aligned} &\textbf{(a)} \quad \cos z = \cos x \cosh y - i \sin x \sinh y \\ &\textbf{(b)} \quad \sin z = \sin x \cosh y + i \cos x \sinh y \end{aligned}$$

and

$$(7) \qquad \begin{aligned} &\textbf{(a)} \quad |\cos z|^2 = \cos^2 x + \sinh^2 y \\ &\textbf{(b)} \quad |\sin z|^2 = \sin^2 x + \sinh^2 y \end{aligned}$$

and give some applications of these formulas.

Solution. From (1),

$$\begin{aligned} \cos z &= \tfrac{1}{2}(e^{i(x+iy)} + e^{-i(x+iy)}) \\ &= \tfrac{1}{2}e^{-y}(\cos x + i \sin x) + \tfrac{1}{2}e^{y}(\cos x - i \sin x) \\ &= \tfrac{1}{2}(e^{y} + e^{-y}) \cos x - \tfrac{1}{2}i(e^{y} - e^{-y}) \sin x. \end{aligned}$$

This yields (6a) since, as is known from calculus,

$$(8) \qquad \cosh y = \tfrac{1}{2}(e^{y} + e^{-y}), \qquad \sinh y = \tfrac{1}{2}(e^{y} - e^{-y});$$

(6b) is obtained similarly. From (6a) and $\cosh^2 y = 1 + \sinh^2 y$ we obtain

$$|\cos z|^2 = (\cos^2 x)(1 + \sinh^2 y) + \sin^2 x \sinh^2 y.$$

Since $\sin^2 x + \cos^2 x = 1$, this gives (7a), and (7b) is obtained similarly.

For instance, $\cos(2 + 3i) = \cos 2 \cosh 3 - i \sin 2 \sinh 3 = -4.190 - 9.109i$.

From (6) we see that $\sin z$ and $\cos z$ are ***periodic with period*** 2π, just as in real. Periodicity of $\tan z$ and $\cot z$ with period π now follows.

Formula (7) points to an essential difference between the real and the complex cosine and sine; whereas $|\cos x| \leqq 1$ and $|\sin x| \leqq 1$, the complex cosine and sine functions are ***no longer bounded*** but approach infinity in absolute value as $y \to \infty$, since then $\sinh y \to \infty$ in (7). ■

EXAMPLE 2 Solutions of Equations. Zeros of cos z and sin z

Solve (a) $\cos z = 5$ (which has no real solution!), (b) $\cos z = 0$, (c) $\sin z = 0$.

Solution. (a) $e^{2iz} - 10e^{iz} + 1 = 0$ from (1) by multiplication by e^{iz}. This is a quadratic equation in e^{iz}, with solutions (rounded off to 3 decimals)

$$e^{iz} = e^{-y+ix} = 5 \pm \sqrt{25 - 1} = 9.899 \quad \text{and} \quad 0.101.$$

Thus $e^{-y} = 9.899$ or 0.101, $e^{ix} = 1$, $y = \pm 2.292$, $x = 2n\pi$. *Ans.* $z = \pm 2n\pi \pm 2.292i$ $(n = 0, 1, 2, \cdots)$.

Can you obtain this from (6a)?

(b) $\cos x = 0$, $\sinh y = 0$ by (7a), $y = 0$. *Ans.* $z = \pm\frac{1}{2}(2n + 1)\pi$ $(n = 0, 1, 2, \cdots)$.
(c) $\sin x = 0$, $\sinh y = 0$ by (7b), *Ans.* $z = \pm n\pi$ $(n = 0, 1, 2, \cdots)$.
Hence the only zeros of $\cos z$ and $\sin z$ are those of the real cosine and sine functions. ■

General formulas *for the real trigonometric functions continue to hold for complex values.* This follows immediately from the definitions. We mention in particular the addition rules

$$\begin{aligned} \cos(z_1 \pm z_2) &= \cos z_1 \cos z_2 \mp \sin z_1 \sin z_2 \\ \sin(z_1 \pm z_2) &= \sin z_1 \cos z_2 \pm \sin z_2 \cos z_1 \end{aligned} \tag{9}$$

and the formula

$$\cos^2 z + \sin^2 z = 1. \tag{10}$$

Some further useful formulas are included in the problem set.

Hyperbolic Functions

The complex **hyperbolic cosine** and **sine** are defined by the formulas

$$\cosh z = \tfrac{1}{2}(e^z + e^{-z}), \qquad \sinh z = \tfrac{1}{2}(e^z - e^{-z}). \tag{11}$$

This is suggested by the familiar definitions for a real variable [see (8)]. These functions are entire, with derivatives

$$(\cosh z)' = \sinh z, \qquad (\sinh z)' = \cosh z, \tag{12}$$

as in calculus. The other hyperbolic functions are defined by

$$\begin{aligned} \tanh z &= \frac{\sinh z}{\cosh z}, \qquad \coth z = \frac{\cosh z}{\sinh z}, \\ \operatorname{sech} z &= \frac{1}{\cosh z}, \qquad \operatorname{csch} z = \frac{1}{\sinh z}. \end{aligned} \tag{13}$$

Complex Trigonometric and Hyperbolic Functions Are Related. If in (11), we replace z by iz and then use (1), we obtain

$$\cosh iz = \cos z, \qquad \sinh iz = i \sin z. \tag{14}$$

Similarly, if in (1) we replace z by iz and then use (11), we obtain conversely

$$\cos iz = \cosh z, \qquad \sin iz = i \sinh z. \tag{15}$$

Here we have another case of *unrelated* real functions that have *related* complex analogs, pointing again to the advantage of working in complex in order to get both a more unified formalism and a deeper understanding of special functions. This is one of the main reasons for the importance of complex analysis to the engineer and physicist.

PROBLEM SET 13.6

1–4 FORMULAS FOR HYPERBOLIC FUNCTIONS

Show that

1. $\cosh z = \cosh x \cos y + i \sinh x \sin y$

   $\sinh z = \sinh x \cos y + i \cosh x \sin y.$

2. $\cosh (z_1 + z_2) = \cosh z_1 \cosh z_2 + \sinh z_1 \sinh z_2$

   $\sinh (z_1 + z_2) = \sinh z_1 \cosh z_2 + \cosh z_1 \sinh z_2.$

3. $\cosh^2 z - \sinh^2 z = 1, \quad \cosh^2 z + \sinh^2 z = \cosh 2z$

4. **Entire Functions.** Prove that $\cos z$, $\sin z$, $\cosh z$, and $\sinh z$ are entire.

5. **Harmonic Functions.** Verify by differentiation that $\operatorname{Im} \cos z$ and $\operatorname{Re} \sin z$ are harmonic.

6–12 Function Values. Find, in the form $u + iv$,

6. $\sin 2\pi i$
7. $\cos i, \quad \sin i$
8. $\cos \pi i, \quad \cosh \pi i$
9. $\cosh (-1 + 2i), \quad \cos (-2 - i)$
10. $\sinh (3 + 4i), \quad \cosh (3 + 4i)$
11. $\sin \pi i, \quad \cos (\frac{1}{2}\pi - \pi i)$
12. $\cos \frac{1}{2}\pi i, \quad \cos [\frac{1}{2}\pi(1 + i)]$

13–15 Equations and Inequalities. Using the definitions, prove:

13. $\cos z$ is even, $\cos (-z) = \cos z$, and $\sin z$ is odd, $\sin (-z) = -\sin z$.

14. $|\sinh y| \leqq |\cos z| \leqq \cosh y, |\sinh y| \leqq |\sin z| \leqq \cosh y$. Conclude that the complex cosine and sine are not bounded in the whole complex plane.

15. $\sin z_1 \cos z_2 = \frac{1}{2}[\sin(z_1 + z_2) + \sin(z_1 - z_2)]$

16–19 Equations. Find all solutions.

16. $\sin z = 100$
17. $\cosh z = 0$
18. $\cosh z = -1$
19. $\sinh z = 0$

20. **Re tan *z* and Im tan *z*.** Show that

$$\operatorname{Re} \tan z = \frac{\sin x \cos x}{\cos^2 x + \sinh^2 y},$$

$$\operatorname{Im} \tan z = \frac{\sinh y \cosh y}{\cos^2 x + \sinh^2 y}.$$

13.7 Logarithm. General Power. Principal Value

We finally introduce the *complex logarithm*, which is more complicated than the real logarithm (which it includes as a special case) and historically puzzled mathematicians for some time (so if you first get puzzled—which need not happen!—be patient and work through this section with extra care).

The **natural logarithm** of $z = x + iy$ is denoted by $\ln z$ (sometimes also by $\log z$) and is defined as the inverse of the exponential function; that is, $w = \ln z$ is defined for $z \neq 0$ by the relation

$$e^w = z.$$

(Note that $z = 0$ is impossible, since $e^w \neq 0$ for all w; see Sec. 13.5.) If we set $w = u + iv$ and $z = re^{i\theta}$, this becomes

$$e^w = e^{u+iv} = re^{i\theta}.$$

Now, from Sec. 13.5, we know that e^{u+iv} has the absolute value e^u and the argument v. These must be equal to the absolute value and argument on the right:

$$e^u = r, \qquad v = \theta.$$

$e^u = r$ gives $u = \ln r$, where $\ln r$ is the familiar *real* natural logarithm of the positive number $r = |z|$. Hence $w = u + iv = \ln z$ is given by

$$\ln z = \ln r + i\theta \qquad (r = |z| > 0, \quad \theta = \arg z). \tag{1}$$

Now comes an important point (without analog in real calculus). Since the argument of z is determined only up to integer multiples of 2π, ***the complex natural logarithm*** **$\ln z$ $(z \neq 0)$** ***is infinitely many-valued.***

The value of $\ln z$ corresponding to the principal value $\operatorname{Arg} z$ (see Sec. 13.2) is denoted by $\operatorname{Ln} z$ (Ln with capital L) and is called the **principal value** of $\ln z$. Thus

$$\operatorname{Ln} z = \ln |z| + i \operatorname{Arg} z \qquad (z \neq 0). \tag{2}$$

The uniqueness of $\operatorname{Arg} z$ for given z $(\neq 0)$ implies that $\operatorname{Ln} z$ is single-valued, that is, a function in the usual sense. Since the other values of $\arg z$ differ by integer multiples of 2π, the other values of $\ln z$ are given by

$$\ln z = \operatorname{Ln} z \pm 2n\pi i \qquad (n = 1, 2, \cdots). \tag{3}$$

They all have the same real part, and their imaginary parts differ by integer multiples of 2π.

If z is positive real, then $\operatorname{Arg} z = 0$, and $\operatorname{Ln} z$ becomes identical with the real natural logarithm known from calculus. If z is negative real (so that the natural logarithm of calculus is not defined!), then $\operatorname{Arg} z = \pi$ and

$$\operatorname{Ln} z = \ln |z| + \pi i \qquad (z \text{ negative real}).$$

From (1) and $e^{\ln r} = r$ for positive real r we obtain

$$e^{\ln z} = z \tag{4a}$$

as expected, but since $\arg (e^z) = y \pm 2n\pi$ is multivalued, so is

$$\ln (e^z) = z \pm 2n\pi i, \qquad n = 0, 1, \cdots. \tag{4b}$$

EXAMPLE 1 Natural Logarithm. Principal Value

$$\begin{aligned}
\ln 1 &= 0, \pm 2\pi i, \pm 4\pi i, \cdots & \operatorname{Ln} 1 &= 0\\
\ln 4 &= 1.386294 \pm 2n\pi i & \operatorname{Ln} 4 &= 1.386294\\
\ln (-1) &= \pm\pi i, \pm 3\pi i, \pm 5\pi i, \cdots & \operatorname{Ln} (-1) &= \pi i\\
\ln (-4) &= 1.386294 \pm (2n + 1)\pi i & \operatorname{Ln} (-4) &= 1.386294 + \pi i\\
\ln i &= \pi i/2, -3\pi/2, 5\pi i/2, \cdots & \operatorname{Ln} i &= \pi i/2\\
\ln 4i &= 1.386294 + \pi i/2 \pm 2n\pi i & \operatorname{Ln} 4i &= 1.386294 + \pi i/2\\
\ln (-4i) &= 1.386294 - \pi i/2 \pm 2n\pi i & \operatorname{Ln} (-4i) &= 1.386294 - \pi i/2\\
\ln (3 - 4i) &= \ln 5 + i \arg (3 - 4i) & \operatorname{Ln} (3 - 4i) &= 1.609438 - 0.927295i\\
&= 1.609438 - 0.927295i \pm 2n\pi i & &\text{(Fig. 337)}
\end{aligned}$$

■

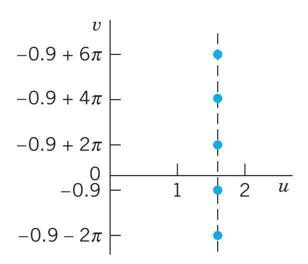


Fig. 337. Some values of ln (3 − 4*i*) in Example 1

The familiar relations for the natural logarithm continue to hold for complex values, that is,

$$\text{(5)} \qquad \text{(a)} \quad \ln(z_1 z_2) = \ln z_1 + \ln z_2, \qquad \text{(b)} \quad \ln(z_1/z_2) = \ln z_1 - \ln z_2$$

but these relations are to be understood in the sense that each value of one side is also contained among the values of the other side; see the next example.

EXAMPLE 2 Illustration of the Functional Relation (5) in Complex

Let

$$z_1 = z_2 = e^{\pi i} = -1.$$

If we take the principal values

$$\text{Ln}\, z_1 = \text{Ln}\, z_2 = \pi i,$$

then (5a) holds provided we write $\ln(z_1 z_2) = \ln 1 = 2\pi i$; however, it is not true for the principal value, $\text{Ln}(z_1 z_2) = \text{Ln}\, 1 = 0$. ■

THEOREM 1 Analyticity of the Logarithm

For every $n = 0, \pm 1, \pm 2, \cdots$ *formula* (3) *defines a function, which is analytic, except at* 0 *and on the negative real axis, and has the derivative*

$$\text{(6)} \qquad (\ln z)' = \frac{1}{z} \qquad (z \text{ not } 0 \text{ or negative real}).$$

PROOF We show that the Cauchy–Riemann equations are satisfied. From (1)–(3) we have

$$\ln z = \ln r + i(\theta + c) = \frac{1}{2}\ln(x^2 + y^2) + i\left(\arctan\frac{y}{x} + c\right)$$

where the constant c is a multiple of 2π. By differentiation,

$$u_x = \frac{x}{x^2 + y^2} = v_y = \frac{1}{1 + (y/x)^2} \cdot \frac{1}{x}$$

$$u_y = \frac{y}{x^2 + y^2} = -v_x = -\frac{1}{1 + (y/x)^2}\left(-\frac{y}{x^2}\right).$$

Hence the Cauchy–Riemann equations hold. [Confirm this by using these equations in polar form, which we did not use since we proved them only in the problems (to Sec. 13.4).] Formula (4) in Sec. 13.4 now gives (6),

$$(\ln z)' = u_x + iv_x = \frac{x}{x^2 + y^2} + i\frac{1}{1 + (y/x)^2}\left(-\frac{y}{x^2}\right) = \frac{x - iy}{x^2 + y^2} = \frac{1}{z}. \qquad \blacksquare$$

Each of the infinitely many functions in (3) is called a **branch** of the logarithm. The negative real axis is known as a **branch cut** and is usually graphed as shown in Fig. 338. The branch for $n = 0$ is called the **principal branch** of $\ln z$.

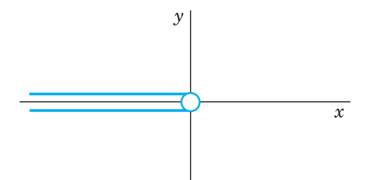


Fig. 338. Branch cut for ln z

General Powers

General powers of a complex number $z = x + iy$ are defined by the formula

$$\textbf{(7)} \qquad z^c = e^{c \ln z} \qquad (c \text{ complex}, z \neq 0).$$

Since $\ln z$ is infinitely many-valued, z^c will, in general, be multivalued. The particular value

$$z^c = e^{c \operatorname{Ln} z}$$

is called the **principal value** *of* z^c.

If $c = n = 1, 2, \cdots$, then z^n is single-valued and identical with the usual nth power of z. If $c = -1, -2, \cdots$, the situation is similar.

If $c = 1/n$, where $n = 2, 3, \cdots$, then

$$z^c = \sqrt[n]{z} = e^{(1/n) \ln z} \qquad (z \neq 0),$$

the exponent is determined up to multiples of $2\pi i/n$ and we obtain the n distinct values of the nth root, in agreement with the result in Sec. 13.2. If $c = p/q$, the quotient of two positive integers, the situation is similar, and z^c has only finitely many distinct values. However, if c is real irrational or genuinely complex, then z^c is infinitely many-valued.

EXAMPLE 3 General Power

$$i^i = e^{i \ln i} = \exp(i \ln i) = \exp\left[i\left(\frac{\pi}{2}i \pm 2n\pi i\right)\right] = e^{-(\pi/2)\mp 2n\pi}.$$

All these values are real, and the principal value ($n = 0$) is $e^{-\pi/2}$.

Similarly, by direct calculation and multiplying out in the exponent,

$$\begin{aligned}(1 + i)^{2-i} &= \exp[(2 - i)\ln(1 + i)] = \exp[(2 - i)\{\ln\sqrt{2} + \tfrac{1}{4}\pi i \pm 2n\pi i\}] \\ &= 2e^{\pi/4 \pm 2n\pi}[\sin(\tfrac{1}{2}\ln 2) + i\cos(\tfrac{1}{2}\ln 2)].\end{aligned} \qquad \blacksquare$$

It is a ***convention*** that for real positive $z = x$ the expression z^c means $e^{c \ln x}$ where $\ln x$ is the elementary real natural logarithm (that is, the principal value $\text{Ln } z$ $(z = x > 0)$ in the sense of our definition). Also, if $z = e$, the base of the natural logarithm, $z^c = e^c$ is *conventionally* regarded as the unique value obtained from (1) in Sec. 13.5.

From (7) we see that for any complex number a,

$$a^z = e^{z \ln a}. \tag{8}$$

We have now introduced the complex functions needed in practical work, some of them (e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$) entire (Sec. 13.5), some of them ($\tan z$, $\cot z$, $\tanh z$, $\coth z$) analytic except at certain points, and one of them ($\ln z$) splitting up into infinitely many functions, each analytic except at 0 and on the negative real axis.

For the **inverse trigonometric** and **hyperbolic functions** see the problem set.

PROBLEM SET 13.7

1–4 VERIFICATIONS IN THE TEXT

1. Verify the computations in Example 1.
2. Verify (5) for $z_1 = -i$ and $z_2 = -1$.
3. Prove analyticity of $\text{Ln } z$ by means of the Cauchy–Riemann equations in polar form (Sec. 13.4).
4. Prove (4a) and (4b).

COMPLEX NATURAL LOGARITHM *ln z*

5–11 Principal Value Ln z. Find $\text{Ln } z$ when z equals

5. -11
6. $4 + 4i$
7. $4 - 4i$
8. $1 \pm i$
9. $0.6 + 0.8i$
10. $-15 \pm 0.1i$
11. ei

12–16 All Values of ln z. Find all values and graph some of them in the complex plane.

12. $\ln e$
13. $\ln 1$
14. $\ln(-7)$
15. $\ln(e^i)$
16. $\ln(4 + 3i)$
17. Show that the set of values of $\ln(i^2)$ differs from the set of values of $2 \ln i$.

18–21 Equations. Solve for z.

18. $\ln z = -\pi i/2$
19. $\ln z = 4 - 3i$
20. $\ln z = e - \pi i$
21. $\ln z = 0.6 + 0.4i$

22–28 General Powers. Find the principal value. Show details.

22. $(2i)^{2i}$
23. $(1 + i)^{1-i}$
24. $(1 - i)^{1+i}$
25. $(-3)^{3-i}$
26. $(i)^{i/2}$
27. $(-1)^{2-i}$
28. $(3 + 4i)^{1/3}$
29. How can you find the answer to Prob. 24 from the answer to Prob. 23?
30. **TEAM PROJECT. Inverse Trigonometric and Hyperbolic Functions.** By definition, the **inverse sine** $w = \arcsin z$ is the relation such that $\sin w = z$. The **inverse cosine** $w = \arccos z$ is the relation such that $\cos w = z$. The **inverse tangent**, **inverse cotangent**, **inverse hyperbolic sine**, etc., are defined and denoted in a similar fashion. (Note that all these relations are ***multivalued***.) Using $\sin w = (e^{iw} - e^{-iw})/(2i)$ and similar representations of $\cos w$, etc., show that

(a) $\arccos z = -i \ln(z + \sqrt{z^2 - 1})$

(b) $\arcsin z = -i \ln(iz + \sqrt{1 - z^2})$

(c) $\text{arccosh } z = \ln(z + \sqrt{z^2 - 1})$

(d) $\text{arcsinh } z = \ln(z + \sqrt{z^2 + 1})$

(e) $\arctan z = \dfrac{i}{2} \ln \dfrac{i + z}{i - z}$

(f) $\text{arctanh } z = \dfrac{1}{2} \ln \dfrac{1 + z}{1 - z}$

(g) Show that $w = \arcsin z$ is infinitely many-valued, and if w_1 is one of these values, the others are of the form $w_1 \pm 2n\pi$ and $\pi - w_1 \pm 2n\pi$, $n = 0, 1, \cdots$. (The *principal value of* $w = u + iv = \arcsin z$ is defined to be the value for which $-\pi/2 \leqq u \leqq \pi/2$ if $v \geqq 0$ and $-\pi/2 < u < \pi/2$ if $v < 0$.)

CHAPTER 13 REVIEW QUESTIONS AND PROBLEMS

1. Divide $15 + 23i$ by $-3 + 7i$. Check the result by multiplication.

2. What happens to a quotient if you take the complex conjugates of the two numbers? If you take the absolute values of the numbers?

3. Write the two numbers in Prob. 1 in polar form. Find the principal values of their arguments.

4. State the definition of the derivative from memory. Explain the big difference from that in calculus.

5. What is an analytic function of a complex variable?

6. Can a function be differentiable at a point without being analytic there? If yes, give an example.

7. State the Cauchy–Riemann equations. Why are they of basic importance?

8. Discuss how e^z, $\cos z$, $\sin z$, $\cosh z$, $\sinh z$ are related.

9. $\ln z$ is more complicated than $\ln x$. Explain. Give examples.

10. How are general powers defined? Give an example. Convert it to the form $x + iy$.

11–16 **Complex Numbers.** Find, in the form $x + iy$, showing details,

11. $(2 + 3i)^2$ **12.** $(1 - i)^{10}$

13. $1/(4 + 3i)$ **14.** $\sqrt{i}$

15. $(1 + i)/(1 - i)$ **16.** $e^{\pi i/2}$, $e^{-\pi i/2}$

17–20 **Polar Form.** Represent in polar form, with the principal argument.

17. $-4 - 4i$ **18.** $12 + i$, $12 - i$

19. $-15i$ **20.** $0.6 + 0.8i$

21–24 **Roots.** Find and graph all values of:

21. $\sqrt{81}$ **22.** $\sqrt{-32i}$

23. $\sqrt[4]{-1}$ **24.** $\sqrt[3]{1}$

25–30 **Analytic Functions.** Find $f(z) = u(x, y) + iv(x, y)$ with u or v as given. Check by the Cauchy–Riemann equations for analyticity.

25 $u = xy$ **26.** $v = y/(x^2 + y^2)$

27. $v = -e^{-2x} \sin 2y$ **28.** $u = \cos 3x \cosh 3y$

29. $u = \exp(-(x^2 - y^2)/2) \cos xy$

30. $v = \cos 2x \sinh 2y$

31–35 **Special Function Values.** Find the value of:

31. $\cos(3 - i)$ **32.** $\text{Ln}(0.6 + 0.8i)$

33. $\tan i$

34. $\sinh(1 + \pi i)$, $\sin(1 + \pi i)$

35. $\cosh(\pi + \pi i)$

SUMMARY OF CHAPTER 13
Complex Numbers and Functions. Complex Differentiation

For arithmetic operations with **complex numbers**

$$z = x + iy = re^{i\theta} = r(\cos\theta + i\sin\theta), \tag{1}$$

$r = |z| = \sqrt{x^2 + y^2}$, $\theta = \arctan(y/x)$, and for their representation in the complex plane, see Secs. 13.1 and 13.2.

A complex function $f(z) = u(x, y) + iv(x, y)$ is **analytic** in a domain D if it has a **derivative** (Sec. 13.3)

$$f'(z) = \lim_{\Delta z \to 0} \frac{f(z + \Delta z) - f(z)}{\Delta z} \tag{2}$$

everywhere in D. Also, $f(z)$ is *analytic at a point* $z = z_0$ if it has a derivative in a neighborhood of z_0 (not merely at z_0 itself).

If $f(z)$ is analytic in D, then $u(x, y)$ and $v(x, y)$ satisfy the (very important!) **Cauchy–Riemann equations** (Sec. 13.4)

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x} \tag{3}$$

everywhere in D. Then u and v also satisfy **Laplace's equation**

$$u_{xx} + u_{yy} = 0, \qquad v_{xx} + v_{yy} = 0 \tag{4}$$

everywhere in D. If $u(x, y)$ and $v(x, y)$ are continuous and have *continuous* partial derivatives in D that satisfy (3) in D, then $f(z) = u(x, y) + iv(x, y)$ is analytic in D. See Sec. 13.4. (More on Laplace's equation and complex analysis follows in Chap. 18.)

The complex **exponential function** (Sec. 13.5)

$$e^z = \exp z = e^x (\cos y + i \sin y) \tag{5}$$

reduces to e^x if $z = x$ $(y = 0)$. It is periodic with $2\pi i$ and has the derivative e^z.

The **trigonometric functions** are (Sec. 13.6)

$$\begin{aligned} \cos z &= \tfrac{1}{2}(e^{iz} + e^{-iz}) = \cos x \cosh y - i \sin x \sinh y \\ \sin z &= \frac{1}{2i}(e^{iz} - e^{-iz}) = \sin x \cosh y + i \cos x \sinh y \end{aligned} \tag{6}$$

and, furthermore,

$$\tan z = (\sin z)/\cos z, \qquad \cot z = 1/\tan z, \quad \text{etc.}$$

The **hyperbolic functions** are (Sec. 13.6)

$$\cosh z = \tfrac{1}{2}(e^z + e^{-z}) = \cos iz, \qquad \sinh z = \tfrac{1}{2}(e^z - e^{-z}) = -i \sin iz \tag{7}$$

etc. The functions (5)–(7) are **entire**, that is, analytic everywhere in the complex plane.

The **natural logarithm** is (Sec. 13.7)

$$\ln z = \ln|z| + i \arg z = \ln|z| + i \operatorname{Arg} z \pm 2n\pi i \tag{8}$$

where $z \neq 0$ and $n = 0, 1, \cdots$. Arg z is the **principal value** of arg z, that is, $-\pi < \operatorname{Arg} z \leqq \pi$. We see that $\ln z$ is infinitely many-valued. Taking $n = 0$ gives the **principal value** Ln z of ln z; thus $\operatorname{Ln} z = \ln|z| + i \operatorname{Arg} z$.

General powers are defined by (Sec. 13.7)

$$z^c = e^{c \ln z} \qquad (c \text{ complex}, z \neq 0). \tag{9}$$

CHAPTER 14

Complex Integration

Chapter 13 laid the groundwork for the study of complex analysis, covered complex numbers in the complex plane, limits, and differentiation, and introduced the most important concept of analyticity. A complex function is *analytic* in some domain if it is differentiable in that domain. Complex analysis deals with such functions and their applications. The Cauchy–Riemann equations, in Sec. 13.4, were the heart of Chapter 13 and allowed a means of checking whether a function is indeed analytic. In that section, we also saw that analytic functions satisfy Laplace's equation, the most important PDE in physics.

We now consider the next part of complex calculus, that is, we shall discuss the first approach to complex integration. It centers around the very important ***Cauchy integral theorem*** (also called the *Cauchy–Goursat theorem*) in Sec. 14.2. This theorem is important because it allows, through its implied ***Cauchy integral formula*** of Sec. 14.3, the evaluation of integrals having an analytic integrand. Furthermore, the Cauchy integral formula shows the surprising result that analytic functions have derivatives of all orders. Hence, in this respect, complex analytic functions behave much more simply than real-valued functions of real variables, which may have derivatives only up to a certain order.

Complex integration is attractive for several reasons. Some basic properties of analytic functions are difficult to prove by other methods. This includes the existence of derivatives of all orders just discussed. A main practical reason for the importance of integration in the complex plane is that such integration can evaluate certain real integrals that appear in applications and that are not accessible by real integral calculus.

Finally, complex integration is used in connection with special functions, such as gamma functions (consult [GenRef1]), the error function, and various polynomials (see [GenRef10]). These functions are applied to problems in physics.

The second approach to complex integration is integration by residues, which we shall cover in Chapter 16.

Prerequisite: Chap. 13.
Section that may be omitted in a shorter course: 14.1, 14.5.
References and Answers to Problems: App. 1 Part D, App. 2.

14.1 Line Integral in the Complex Plane

As in calculus, in complex analysis we distinguish between definite integrals and indefinite integrals or antiderivatives. Here an **indefinite integral** is a function whose derivative equals a given analytic function in a region. By inverting known differentiation formulas we may find many types of indefinite integrals.

Complex definite integrals are called (complex) **line integrals**. They are written

$$\int_C f(z)\, dz.$$

Here the **integrand** $f(z)$ is integrated over a given curve C or a portion of it (an *arc*, but we shall say "***curve***" in either case, for simplicity). This curve C in the complex plane is called the **path of integration**. We may represent C by a parametric representation

$$z(t) = x(t) = iy(t) \qquad (a \leqq t \leqq b). \tag{1}$$

The sense of increasing t is called the **positive sense** on C, and we say that C is **oriented** by (1).

For instance, $z(t) = t + 3it$ $(0 \leqq t \leqq 2)$ gives a portion (a segment) of the line $y = 3x$. The function $z(t) = 4 \cos t + 4i \sin t$ $(-\pi \leqq t \leqq \pi)$ represents the circle $|z| = 4$, and so on. More examples follow below.

We assume C to be a **smooth curve**, that is, C has a continuous and nonzero derivative

$$\dot{z}(t) = \frac{dz}{dt} = \dot{x}(t) + i\dot{y}(t)$$

at each point. Geometrically this means that C has everywhere a continuously turning tangent, as follows directly from the definition

$$\dot{z}(t) = \lim_{\Delta t \to 0} \frac{z(t + \Delta t) - z(t)}{\Delta t} \qquad \text{(Fig. 339).}$$

Here we use a dot since a prime $'$ denotes the derivative with respect to z.

Definition of the Complex Line Integral

This is similar to the method in calculus. Let C be a smooth curve in the complex plane given by (1), and let $f(z)$ be a continuous function given (at least) at each point of C. We now subdivide (we "***partition***") the interval $a \leqq t \leqq b$ in (1) by points

$$t_0\,(= a), \quad t_1, \quad \cdots, \quad t_{n-1}, \quad t_n\,(= b)$$

where $t_0 < t_1 < \cdots < t_n$. To this subdivision there corresponds a subdivision of C by points

$$z_0, \quad z_1, \quad \cdots, \quad z_{n-1}, \quad z_n\,(= Z) \qquad \text{(Fig. 340),}$$

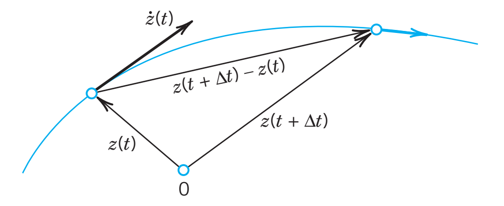


Fig. 339. Tangent vector $\dot{z}(t)$ of a curve C in the complex plane given by $z(t)$. The arrowhead on the curve indicates the *positive sense* (sense of increasing t)

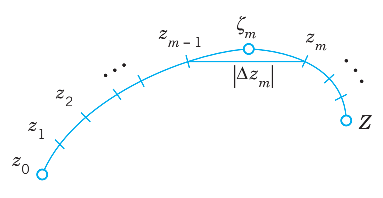


Fig. 340. Complex line integral

where $z_j = z(t_j)$. On each portion of subdivision of C we choose an arbitrary point, say, a point ζ_1 between z_0 and z_1 (that is, $\zeta_1 = z(t)$ where t satisfies $t_0 \leqq t \leqq t_1$), a point ζ_2 between z_1 and z_2, etc. Then we form the sum

$$S_n = \sum_{m=1}^{n} f(\zeta_m)\, \Delta z_m \qquad \text{where} \qquad \Delta z_m = z_m - z_{m-1}. \tag{2}$$

We do this for each $n = 2, 3, \cdots$ in a completely independent manner, but so that the greatest $|\Delta t_m| = |t_m - t_{m-1}|$ approaches zero as $n \to \infty$. This implies that the greatest $|\Delta z_m|$ also approaches zero. Indeed, it cannot exceed the length of the arc of C from z_{m-1} to z_m and the latter goes to zero since the arc length of the smooth curve C is a continuous function of t. The limit of the sequence of complex numbers $S_2, S_3, \cdots$ thus obtained is called the **line integral** (or simply the *integral*) of $f(z)$ over the path of integration C with the orientation given by (1). This line integral is denoted by

$$\int_C f(z)\, dz, \qquad \text{or by} \qquad \oint_C f(z)\, dz \tag{3}$$

if C is a **closed path** (one whose terminal point Z coincides with its initial point z_0, as for a circle or for a curve shaped like an 8).

General Assumption. *All paths of integration for complex line integrals are assumed to be* **piecewise smooth**, *that is, they consist of finitely many smooth curves joined end to end.*

Basic Properties Directly Implied by the Definition

1. **Linearity.** Integration is a **linear operation**, that is, we can integrate sums term by term and can take out constant factors from under the integral sign. This means that if the integrals of f_1 and f_2 over a path C exist, so does the integral of $k_1 f_1 + k_2 f_2$ over the same path and

$$\int_C [k_1 f_1(z) + k_2 f_2(z)]\, dz = k_1 \int_C f_1(z)\, dz + k_2 \int_C f_2(z)\, dz. \tag{4}$$

2. **Sense reversal** in integrating over the ***same*** path, from z_0 to Z (left) and from Z to z_0 (right), introduces a minus sign as shown,

$$\int_{z_0}^{Z} f(z)\, dz = -\int_{Z}^{z_0} f(z)\, dz. \tag{5}$$

3. **Partitioning of path** (see Fig. 341)

$$\int_C f(z)\, dz = \int_{C_1} f(z)\, dz + \int_{C_2} f(z)\, dz. \tag{6}$$

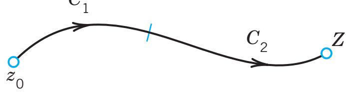


Fig. 341. Partitioning of path [formula (6)]

Existence of the Complex Line Integral

Our assumptions that $f(z)$ is continuous and C is piecewise smooth imply the existence of the line integral (3). This can be seen as follows.

As in the preceding chapter let us write $f(z) = u(x, y) + iv(x, y)$. We also set

$$\zeta_m = \xi_m + i\eta_m \qquad \text{and} \qquad \Delta z_m = \Delta x_m + i\Delta y_m.$$

Then (2) may be written

$$S_n = \sum (u + iv)(\Delta x_m + i\Delta y_m) \tag{7}$$

where $u = u(\zeta_m, \eta_m)$, $v = v(\zeta_m, \eta_m)$ and we sum over m from 1 to n. Performing the multiplication, we may now split up S_n into four sums:

$$S_n = \sum u\,\Delta x_m - \sum v\,\Delta y_m + i\left[\sum u\,\Delta y_m + \sum v\,\Delta x_m\right].$$

These sums are real. Since f is continuous, u and v are continuous. Hence, if we let n approach infinity in the aforementioned way, then the greatest Δx_m and Δy_m will approach zero and each sum on the right becomes a real line integral:

$$\begin{aligned} \lim_{n\to\infty} S_n &= \int_C f(z)\,dz \\ &= \int_C u\,dx - \int_C v\,dy + i\left[\int_C u\,dy + \int_C v\,dx\right]. \end{aligned} \tag{8}$$

This shows that under our assumptions on f and C the line integral (3) exists and its value is independent of the choice of subdivisions and intermediate points ζ_m. ■

First Evaluation Method: Indefinite Integration and Substitution of Limits

This method is the analog of the evaluation of definite integrals in calculus by the well-known formula

$$\int_a^b f(x)\,dx = F(b) - F(a)$$

where $[F'(x) = f(x)]$.

It is simpler than the next method, but it is suitable for analytic functions only. To formulate it, we need the following concept of general interest.

A domain D is called **simply connected** if every **simple closed curve** (closed curve without self-intersections) encloses only points of D.

For instance, a circular disk is simply connected, whereas an annulus (Sec. 13.3) is not simply connected. (Explain!)

THEOREM 1

Indefinite Integration of Analytic Functions

Let $f(z)$ be analytic in a simply connected domain D. Then there exists an indefinite integral of $f(z)$ in the domain D, that is, an analytic function $F(z)$ such that $F'(z) = f(z)$ in D, and for all paths in D joining two points z_0 and z_1 in D we have

$$\int_{z_0}^{z_1} f(z)\, dz = F(z_1) - F(z_0) \qquad [F'(z) = f(z)]. \tag{9}$$

(Note that we can write z_0 and z_1 instead of C, since we get the same value for all those C from z_0 to z_1.)

This theorem will be proved in the next section.

Simple connectedness is quite essential in Theorem 1, as we shall see in Example 5.

Since analytic functions are our main concern, and since differentiation formulas will often help in finding $F(z)$ for a given $f(z) = F'(z)$, the present method is of great practical interest.

If $f(z)$ is entire (Sec. 13.5), we can take for D the complex plane (which is certainly simply connected).

EXAMPLE 1

$$\int_0^{1+i} z^2\, dz = \frac{1}{3} z^3 \Big|_0^{1+i} = \frac{1}{3}(1+i)^3 = -\frac{2}{3} + \frac{2}{3} i$$ ∎

EXAMPLE 2

$$\int_{-\pi i}^{\pi i} \cos z\, dz = \sin z \Big|_{-\pi i}^{\pi i} = 2 \sin \pi i = 2i \sinh \pi = 23.097i$$ ∎

EXAMPLE 3

$$\int_{8+\pi i}^{8-3\pi i} e^{z/2}\, dz = 2e^{z/2} \Big|_{8+\pi i}^{8-3\pi i} = 2(e^{4-3\pi i/2} - e^{4+\pi i/2}) = 0$$

since e^z is periodic with period $2\pi i$. ∎

EXAMPLE 4

$\displaystyle\int_{-i}^{i} \frac{dz}{z} = \operatorname{Ln} i - \operatorname{Ln}(-i) = \frac{i\pi}{2} - \left(-\frac{i\pi}{2}\right) = i\pi$. Here D is the complex plane without 0 and the negative real axis (where $\operatorname{Ln} z$ is not analytic). Obviously, D is a simply connected domain. ∎

Second Evaluation Method: Use of a Representation of a Path

This method is not restricted to analytic functions but applies to any continuous complex function.

THEOREM 2

Integration by the Use of the Path

Let C be a piecewise smooth path, represented by $z = z(t)$, where $a \leqq t \leqq b$. Let $f(z)$ be a continuous function on C. Then

$$\int_C f(z)\, dz = \int_a^b f[z(t)]\dot{z}(t)\, dt \qquad \left(\dot{z} = \frac{dz}{dt}\right). \tag{10}$$

PROOF The left side of (10) is given by (8) in terms of real line integrals, and we show that the right side of (10) also equals (8). We have $z = x + iy$, hence $\dot{z} = \dot{x} + i\dot{y}$. We simply write u for $u[x(t), y(t)]$ and v for $v[x(t), y(t)]$. We also have $dx = \dot{x}\,dt$ and $dy = \dot{y}\,dt$. Consequently, in (10)

$$\begin{aligned}\int_a^b f[z(t)]\dot{z}(t)\,dt &= \int_a^b (u + iv)(\dot{x} + i\dot{y})\,dt \\ &= \int_C [u\,dx - v\,dy + i(u\,dy + v\,dx)] \\ &= \int_C (u\,dx - v\,dy) + i\int_C (u\,dy + v\,dx).\end{aligned}$$

COMMENT. In (7) and (8) of the existence proof of the complex line integral we referred to real line integrals. If one wants to avoid this, one can take (10) as a *definition* of the complex line integral.

Steps in Applying Theorem 2

(A) Represent the path C in the form $z(t)$ $(a \leqq t \leqq b)$.

(B) Calculate the derivative $\dot{z}(t) = dz/dt$.

(C) Substitute $z(t)$ for every z in $f(z)$ (hence $x(t)$ for x and $y(t)$ for y).

(D) Integrate $f[z(t)]\dot{z}(t)$ over t from a to b.

EXAMPLE 5 **A Basic Result: Integral of 1/z Around the Unit Circle**

We show that by integrating $1/z$ counterclockwise around the unit circle (the circle of radius 1 and center 0; see Sec. 13.3) we obtain

$$\oint_C \frac{dz}{z} = 2\pi i \qquad (C \text{ the unit circle, counterclockwise}). \tag{11}$$

This is a very important result that we shall need quite often.

Solution. **(A)** We may represent the unit circle C in Fig. 330 of Sec. 13.3 by

$$z(t) = \cos t + i \sin t = e^{it} \qquad (0 \leqq t \leqq 2\pi),$$

so that counterclockwise integration corresponds to an increase of t from 0 to 2π.

(B) Differentiation gives $\dot{z}(t) = ie^{it}$ (chain rule!).

(C) By substitution, $f(z(t)) = 1/z(t) = e^{-it}$.

(D) From (10) we thus obtain the result

$$\oint_C \frac{dz}{z} = \int_0^{2\pi} e^{-it} ie^{it}\,dt = i\int_0^{2\pi} dt = 2\pi i.$$

Check this result by using $z(t) = \cos t + i \sin t$.

Simple connectedness is essential in Theorem 1. Equation (9) in Theorem 1 gives 0 for any closed path because then $z_1 = z_0$, so that $F(z_1) - F(z_0) = 0$. Now $1/z$ is not analytic at $z = 0$. But any *simply connected* domain containing the unit circle must contain $z = 0$, so that Theorem 1 does not apply—it is not enough that $1/z$ is analytic in an annulus, say, $\frac{1}{2} < |z| < \frac{3}{2}$, because an annulus is not simply connected!

EXAMPLE 6 Integral of $1/z^m$ with Integer Power m

Let $f(z) = (z - z_0)^m$ where m is the integer and z_0 a constant. Integrate counterclockwise around the circle C of radius ρ with center at z_0 (Fig. 342).

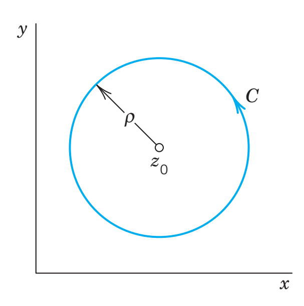


Fig. 342. Path in Example 6

Solution. We may represent C in the form

$$z(t) = z_0 + \rho(\cos t + i \sin t) = z_0 + \rho e^{it} \qquad (0 \leqq t \leqq 2\pi).$$

Then we have

$$(z - z_0)^m = \rho^m e^{imt}, \qquad dz = i\rho e^{it}\, dt$$

and obtain

$$\oint_C (z - z_0)^m\, dz = \int_0^{2\pi} \rho^m e^{imt}\, i\rho e^{it}\, dt = i\rho^{m+1} \int_0^{2\pi} e^{i(m+1)t}\, dt.$$

By the Euler formula (5) in Sec. 13.6 the right side equals

$$i\rho^{m+1}\left[\int_0^{2\pi} \cos(m+1)t\, dt + i\int_0^{2\pi} \sin(m+1)t\, dt\right].$$

If $m = -1$, we have $\rho^{m+1} = 1$, $\cos 0 = 1$, $\sin 0 = 0$. We thus obtain $2\pi i$. For integer $m \neq -1$ each of the two integrals is zero because we integrate over an interval of length 2π, equal to a period of sine and cosine. Hence the result is

$$\oint_C (z - z_0)^m\, dz = \begin{cases} 2\pi i & (m = -1), \\ 0 & (m \neq -1 \text{ and integer}). \end{cases} \tag{12}$$

■

Dependence on path. Now comes a very important fact. If we integrate a given function $f(z)$ from a point z_0 to a point z_1 along different paths, the integrals will in general have different values. In other words, ***a complex line integral depends not only on the endpoints of the path but in general also on the path itself***. The next example gives a first impression of this, and a systematic discussion follows in the next section.

EXAMPLE 7 Integral of a Nonanalytic Function. Dependence on Path

Integrate $f(z) = \text{Re}\, z = x$ from 0 to $1 + 2i$ (a) along C^* in Fig. 343, (b) along C consisting of C_1 and C_2.

Solution. **(a)** C^* can be represented by $z(t) = t + 2it$ $(0 \leqq t \leqq 1)$. Hence $\dot{z}(t) = 1 + 2i$ and $f[z(t)] = x(t) = t$ on C^*. We now calculate

$$\int_{C^*} \text{Re}\, z\, dz = \int_0^1 t(1 + 2i)\, dt = \frac{1}{2}(1 + 2i) = \frac{1}{2} + i.$$

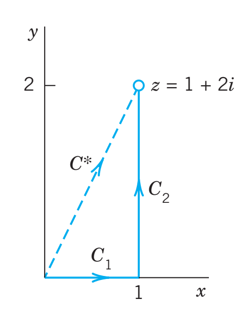


Fig. 343. Paths in Example 7

(b) We now have

$$C_1: z(t) = t, \qquad \dot{z}(t) = 1, \qquad f(z(t)) = x(t) = t \qquad (0 \leqq t \leqq 1)$$

$$C_2: z(t) = 1 + it, \qquad \dot{z}(t) = i, \qquad f(z(t)) = x(t) = 1 \qquad (0 \leqq t \leqq 2).$$

Using (6) we calculate

$$\int_C \operatorname{Re} z\, dz = \int_{C_1} \operatorname{Re} z\, dz + \int_{C_2} \operatorname{Re} z\, dz = \int_0^1 t\, dt + \int_0^2 1 \cdot i\, dt = \frac{1}{2} + 2i.$$

Note that this result differs from the result in (a). ■

Bounds for Integrals. *ML*-Inequality

There will be a frequent need for estimating the absolute value of complex line integrals. The basic formula is

$$\left| \int_C f(z)\, dz \right| \leqq ML \qquad \textbf{(13)} \qquad (\textbf{\textit{ML}-inequality});$$

L is the length of C and M a constant such that $|f(z)| \leqq M$ everywhere on C.

PROOF Taking the absolute value in (2) and applying the generalized inequality (6*) in Sec. 13.2, we obtain

$$|S_n| = \left| \sum_{m=1}^{n} f(\zeta_m)\, \Delta z_m \right| \leqq \sum_{m=1}^{n} |f(\zeta_m)|\,|\Delta z_m| \leqq M \sum_{m=1}^{n} |\Delta z_m|.$$

Now $|\Delta z_m|$ is the length of the chord whose endpoints are z_{m-1} and z_m (see Fig. 340). Hence the sum on the right represents the length L^* of the broken line of chords whose endpoints are $z_0, z_1, \cdots, z_n$ $(= Z)$. If n approaches infinity in such a way that the greatest $|\Delta t_m|$ and thus $|\Delta z_m|$ approach zero, then L^* approaches the length L of the curve C, by the definition of the length of a curve. From this the inequality (13) follows. ■

We cannot see from (13) how close to the bound ML the actual absolute value of the integral is, but this will be no handicap in applying (13). For the time being we explain the practical use of (13) by a simple example.

EXAMPLE 8 **Estimation of an Integral**

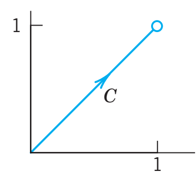


Fig. 344. Path in Example 8

Find an upper bound for the absolute value of the integral

$$\int_C z^2\,dz, \qquad C \text{ the straight-line segment from } 0 \text{ to } 1+i, \text{ Fig. 344.}$$

Solution. $L = \sqrt{2}$ and $|f(z)| = |z^2| \leqq 2$ on C gives by (13)

$$\left|\int_C z^2\,dz\right| \leqq 2\sqrt{2} = 2.8284.$$

The absolute value of the integral is $|-\frac{2}{3} + \frac{2}{3}i| = \frac{2}{3}\sqrt{2} = 0.9428$ (see Example 1). ■

Summary on Integration. Line integrals of $f(z)$ can always be evaluated by (10), using a representation (1) of the path of integration. If $f(z)$ is analytic, indefinite integration by (9) as in calculus will be simpler (proof in the next section).

PROBLEM SET 14.1

1–10 FIND THE PATH and sketch it.

1. $z(t) = (1 + \frac{1}{2}i)t \quad (2 \leqq t \leqq 5)$
2. $z(t) = 3 + i + (1 - i)t \quad (0 \leqq t \leqq 3)$
3. $z(t) = t + 2it^2 \quad (1 \leqq t \leqq 2)$
4. $z(t) = t + (1 - t)^2 i \quad (-1 \leqq t \leqq 1)$
5. $z(t) = 3 - i + \sqrt{10}e^{-it} \quad (0 \leqq t \leqq 2\pi)$
6. $z(t) = 1 + i + e^{-\pi it} \quad (0 \leqq t \leqq 2)$
7. $z(t) = 2 + 4e^{\pi it/2} \quad (0 \leqq t \leqq 2)$
8. $z(t) = 5e^{-it} \quad (0 \leqq t \leqq \pi/2)$
9. $z(t) = t + it^3 \quad (-2 \leqq t \leqq 2)$
10. $z(t) = 2\cos t + i \sin t \quad (0 \leqq t \leqq 2\pi)$

11–20 FIND A PARAMETRIC REPRESENTATION and sketch the path.

11. Segment from $(-1, 1)$ to $(1, 3)$
12. From $(0, 0)$ to $(2, 1)$ along the axes
13. Upper half of $|z - 2 + i| = 2$ from $(4, -1)$ to $(0, -1)$
14. Unit circle, clockwise
15. $x^2 - 4y^2 = 4$, the branch through $(2, 0)$
16. Ellipse $4x^2 + 9y^2 = 36$, counterclockwise
17. $|z + a + ib| = r$, clockwise
18. $y = 1/x$ from $(1, 1)$ to $(5, \frac{1}{5})$
19. Parabola $y = 1 - \frac{1}{4}x^2 \quad (-2 \leqq x \leqq 2)$
20. $4(x - 2)^2 + 5(y + 1)^2 = 20$

21–30 INTEGRATION

Integrate by the first method or state why it does not apply and use the second method. Show the details.

21. $\int_C \text{Re}\, z\,dz$, C the shortest path from $1 + i$ to $3 + 3i$
22. $\int_C \text{Re}\, z\,dz$, C the parabola $y = 1 + \frac{1}{2}(x - 1)^2$ from $1 + i$ to $3 + 3i$
23. $\int_C e^z\,dz$, C the shortest path from πi to $2\pi i$
24. $\int_C \cos 2z\,dz$, C the semicircle $|z| = \pi$, $x \geqq 0$ from $-\pi i$ to πi
25. $\int_C z \exp(z^2)\,dz$, C from 1 along the axes to i
26. $\int_C (z + z^{-1})\,dz$, C the unit circle, counterclockwise
27. $\int_C \sec^2 z\,dz$, any path from $\pi/4$ to $\pi i/4$
28. $\int_C \left(\frac{5}{z - 2i} - \frac{6}{(z - 2i)^2}\right) dz$, C the circle $|z - 2i| = 4$, clockwise
29. $\int_C \text{Im}\, z^2\,dz$ counterclockwise around the triangle with vertices $0, 1, i$
30. $\int_C \text{Re}\, z^2\,dz$ clockwise around the boundary of the square with vertices $0, i, 1 + i, 1$
31. **CAS PROJECT. Integration.** Write programs for the two integration methods. Apply them to problems of your choice. Could you make them into a joint program that also decides which of the two methods to use in a given case?

32. Sense reversal. Verify (5) for $f(z) = z^2$, where C is the segment from $-1 - i$ to $1 + i$.

33. Path partitioning. Verify (6) for $f(z) = 1/z$ and C_1 and C_2 the upper and lower halves of the unit circle.

34. TEAM EXPERIMENT. Integration. (a) Comparison. First write a short report comparing the essential points of the two integration methods.

(b) Comparison. Evaluate $\int_C f(z)\, dz$ by Theorem 1 and check the result by Theorem 2, where:

(i) $f(z) = z^4$ and C is the semicircle $|z| = 2$ from $-2i$ to $2i$ in the right half-plane,

(ii) $f(z) = e^{2z}$ and C is the shortest path from 0 to $1 + 2i$.

(c) Continuous deformation of path. Experiment with a family of paths with common endpoints, say, $z(t) = t + ia \sin t$, $0 \leqq t \leqq \pi$, with real parameter a. Integrate nonanalytic functions (Re z, Re (z^2), etc.) and explore how the result depends on a. Then take analytic functions of your choice. (Show the details of your work.) Compare and comment.

(d) Continuous deformation of path. Choose another family, for example, semi-ellipses $z(t) = a \cos t + i \sin t$, $-\pi/2 \leqq t \leqq \pi/2$, and experiment as in (c).

35. *ML*-inequality. Find an upper bound of the absolute value of the integral in Prob. 21.

14.2 Cauchy's Integral Theorem

This section is the focal point of the chapter. We have just seen in Sec. 14.1 that a line integral of a function $f(z)$ generally depends not merely on the endpoints of the path, but also on the choice of the path itself. This dependence often complicates situations. Hence conditions under which this does ***not*** occur are of considerable importance. Namely, if $f(z)$ is analytic in a domain D and D is simply connected (see Sec. 14.1 and also below), then the integral will not depend on the choice of a path between given points. This result (Theorem 2) follows from Cauchy's integral theorem, along with other basic consequences that make ***Cauchy's integral theorem the most important theorem in this chapter*** and fundamental throughout complex analysis.

Let us continue our discussion of simple connectedness which we started in Sec. 14.1.

1. A **simple closed path** is a closed path (defined in Sec. 14.1) that does not intersect or touch itself as shown in Fig. 345. For example, a circle is simple, but a curve shaped like an 8 is not simple.

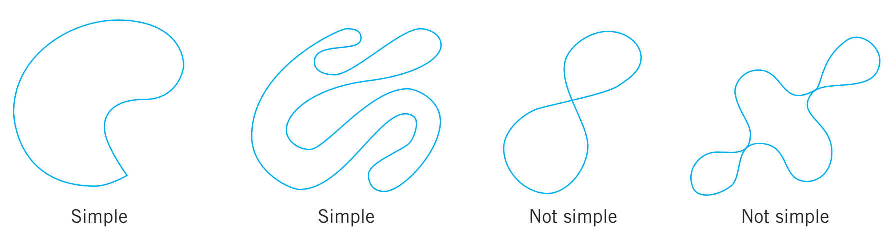


Fig. 345. Closed paths

2. A **simply connected domain** D in the complex plane is a domain (Sec. 13.3) such that every simple closed path in D encloses only points of D. *Examples:* The interior of a circle ("open disk"), ellipse, or any simple closed curve. A domain that is not simply connected is called **multiply connected.** *Examples:* An annulus (Sec. 13.3), a disk without the center, for example, $0 < |z| < 1$. See also Fig. 346.

More precisely, a **bounded domain** D (that is, a domain that lies entirely in some circle about the origin) is called ***p*-fold connected** if its boundary consists of p closed

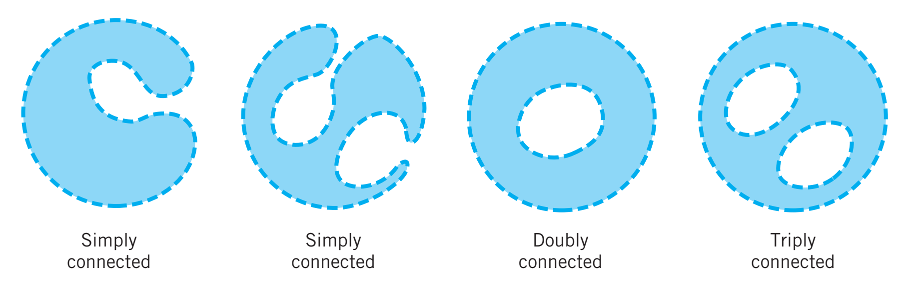


Fig. 346. Simply and multiply connected domains

connected sets without common points. These sets can be curves, segments, or single points (such as $z = 0$ for $0 < |z| < 1$, for which $p = 2$). Thus, D has $p - 1$ "**holes**," where "hole" may also mean a segment or even a single point. Hence an annulus is doubly connected ($p = 2$).

THEOREM 1

Cauchy's Integral Theorem

If $f(z)$ is analytic in a simply connected domain D, then for every simple closed path C in D,

$$\oint_C f(z)\, dz = 0. \qquad (1)$$

See Fig. 347.

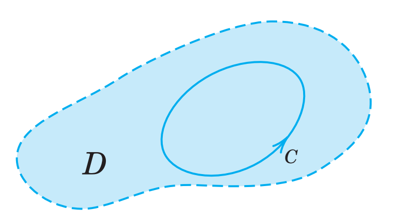


Fig. 347. Cauchy's integral theorem

Before we prove the theorem, let us consider some examples in order to really understand what is going on. A simple closed path is sometimes called a *contour* and an integral over such a path a **contour integral**. Thus, (1) and our examples involve contour integrals.

EXAMPLE 1 **Entire Functions**

$$\oint_C e^z\, dz = 0, \qquad \oint_C \cos z\, dz = 0, \qquad \oint_C z^n\, dz = 0 \qquad (n = 0, 1, \cdots)$$

for any closed path, since these functions are entire (analytic for all z). ■

EXAMPLE 2 **Points Outside the Contour Where $f(x)$ is Not Analytic**

$$\oint_C \sec z\, dz = 0, \qquad \oint_C \frac{dz}{z^2 + 4} = 0$$

where C is the unit circle, $\sec z = 1/\cos z$ is not analytic at $z = \pm\pi/2, \pm 3\pi/2, \cdots$, but all these points lie outside C; none lies on C or inside C. Similarly for the second integral, whose integrand is not analytic at $z = \pm 2i$ outside C. ■

EXAMPLE 3 **Nonanalytic Function**

$$\oint_C \bar{z}\, dz = \int_0^{2\pi} e^{-it} ie^{it}\, dt = 2\pi i$$

where C: $z(t) = e^{it}$ is the unit circle. This does not contradict Cauchy's theorem because $f(z) = \bar{z}$ is not analytic. ∎

EXAMPLE 4 **Analyticity Sufficient, Not Necessary**

$$\oint_C \frac{dz}{z^2} = 0$$

where C is the unit circle. This result does *not* follow from Cauchy's theorem, because $f(z) = 1/z^2$ is not analytic at $z = 0$. Hence *the condition that f be analytic in D is sufficient rather than necessary for* (1) *to be true.* ∎

EXAMPLE 5 **Simple Connectedness Essential**

$$\oint_C \frac{dz}{z} = 2\pi i$$

for counterclockwise integration around the unit circle (see Sec. 14.1). C lies in the annulus $\frac{1}{2} < |z| < \frac{3}{2}$ where $1/z$ is analytic, but this domain is not simply connected, so that Cauchy's theorem cannot be applied. Hence ***the condition that the domain D be simply connected is essential.***

In other words, by Cauchy's theorem, if $f(z)$ is analytic on a simple closed path C and everywhere inside C, with no exception, not even a single point, then (1) holds. The point that causes trouble here is $z = 0$ where $1/z$ is not analytic. ∎

PROOF Cauchy proved his integral theorem under the additional assumption that the derivative $f'(z)$ is continuous (which is true, but would need an extra proof). His proof proceeds as follows. From (8) in Sec. 14.1 we have

$$\oint_C f(z)\, dz = \oint_C (u\, dx - v\, dy) + i \oint_C (u\, dy + v\, dx).$$

Since $f(z)$ is analytic in D, its derivative $f'(z)$ exists in D. Since $f'(z)$ is assumed to be continuous, (4) and (5) in Sec. 13.4 imply that u and v have *continuous* partial derivatives in D. Hence Green's theorem (Sec. 10.4) (with u and $-v$ instead of F_1 and F_2) is applicable and gives

$$\oint_C (u\, dx - v\, dy) = \iint_R \left(-\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y}\right) dx\, dy$$

where R is the region bounded by C. The second Cauchy–Riemann equation (Sec. 13.4) shows that the integrand on the right is identically zero. Hence the integral on the left is zero. In the same fashion it follows by the use of the first Cauchy–Riemann equation that the last integral in the above formula is zero. This completes Cauchy's proof. ∎

Goursat's proof ***without the condition that $f'(z)$ is continuous***[1] is much more complicated. We leave it optional and include it in App. 4.

[1] ÉDOUARD GOURSAT (1858–1936), French mathematician who made important contributions to complex analysis and PDEs. Cauchy published the theorem in 1825. The removal of that condition by Goursat (see *Transactions Amer. Math Soc.*, vol. 1, 1900) is quite important because, for instance, derivatives of analytic functions are also analytic. Because of this, Cauchy's integral theorem is also called Cauchy–Goursat theorem.

Independence of Path

We know from the preceding section that the value of a line integral of a given function $f(z)$ from a point z_1 to a point z_2 will in general depend on the path C over which we integrate, not merely on z_1 and z_2. It is important to characterize situations in which this difficulty of path dependence does not occur. This task suggests the following concept. We call an integral of $f(z)$ **independent of path in a domain *D*** if for every z_1, z_2 in D its value depends (besides on $f(z)$, of course) only on the initial point z_1 and the terminal point z_2, but not on the choice of the path C in D [so that every path in D from z_1 to z_2 gives the same value of the integral of $f(z)$].

THEOREM 2

Independence of Path

If $f(z)$ is analytic in a simply connected domain D, then the integral of $f(z)$ is independent of path in D.

PROOF Let z_1 and z_2 be any points in D. Consider two paths C_1 and C_2 in D from z_1 to z_2 without further common points, as in Fig. 348. Denote by C_2^* the path C_2 with the orientation reversed (Fig. 349). Integrate from z_1 over C_1 to z_2 and over C_2^* back to z_1. This is a simple closed path, and Cauchy's theorem applies under our assumptions of the present theorem and gives zero:

$$(2') \qquad \int_{C_1} f\,dz + \int_{C_2^*} f\,dz = 0, \qquad \text{thus} \qquad \int_{C_1} f\,dz = -\int_{C_2^*} f\,dz.$$

But the minus sign on the right disappears if we integrate in the reverse direction, from z_1 to z_2, which shows that the integrals of $f(z)$ over C_1 and C_2 are equal,

$$(2) \qquad \int_{C_1} f(z)\,dz = \int_{C_2} f(z)\,dz \qquad \text{(Fig. 348).}$$

This proves the theorem for paths that have only the endpoints in common. For paths that have finitely many further common points, apply the present argument to each "loop" (portions of C_1 and C_2 between consecutive common points; four loops in Fig. 350). For paths with infinitely many common points we would need additional argumentation not to be presented here.

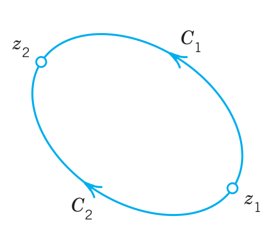


Fig. 348. Formula (2)

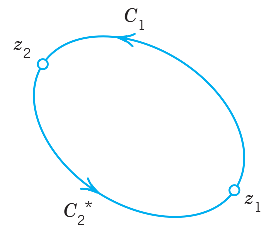


Fig. 349. Formula (2′)

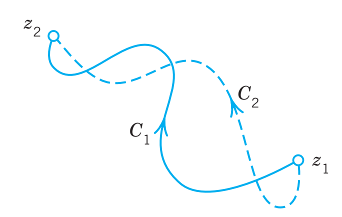


Fig. 350. Paths with more common points

Principle of Deformation of Path

This idea is related to path independence. We may imagine that the path C_2 in (2) was obtained from C_1 by continuously moving C_1 (with ends fixed!) until it coincides with C_2. Figure 351 shows two of the infinitely many intermediate paths for which the integral always retains its value (because of Theorem 2). Hence we may impose a continuous deformation of the path of an integral, keeping the ends fixed. As long as our deforming path always contains only points at which $f(z)$ is analytic, the integral retains the same value. This is called the **principle of deformation of path**.

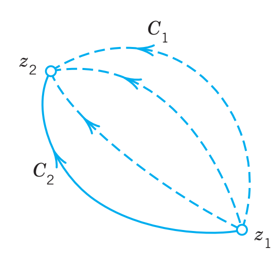


Fig. 351. Continuous deformation of path

EXAMPLE 6 **A Basic Result: Integral of Integer Powers**

From Example 6 in Sec. 14.1 and the principle of deformation of path it follows that

$$\oint (z - z_0)^m \, dz = \begin{cases} 2\pi i & (m = -1) \\ 0 & (m \neq -1 \text{ and integer}) \end{cases} \tag{3}$$

for counterclockwise integration around ***any simple closed path containing z_0 in its interior***.

Indeed, the circle $|z - z_0| = \rho$ in Example 6 of Sec. 14.1 can be continuously deformed in two steps into a path as just indicated, namely, by first deforming, say, one semicircle and then the other one. (Make a sketch). ■

Existence of Indefinite Integral

We shall now justify our indefinite integration method in the preceding section [formula (9) in Sec. 14.1]. The proof will need Cauchy's integral theorem.

THEOREM 3 **Existence of Indefinite Integral**

If $f(z)$ is analytic in a simply connected domain D, then there exists an indefinite integral $F(z)$ of $f(z)$ in D—thus, $F'(z) = f(z)$—which is analytic in D, and for all paths in D joining any two points z_0 and z_1 in D, the integral of $f(z)$ from z_0 to z_1 can be evaluated by formula (9) *in* Sec. 14.1.

PROOF The conditions of Cauchy's integral theorem are satisfied. Hence the line integral of $f(z)$ from any z_0 in D to any z in D is independent of path in D. We keep z_0 fixed. Then this integral becomes a function of z, call if $F(z)$,

$$F(z) = \int_{z_0}^{z} f(z^*) \, dz^* \tag{4}$$

which is uniquely determined. We show that this $F(z)$ is analytic in D and $F'(z) = f(z)$. The idea of doing this is as follows. Using (4) we form the difference quotient

$$\frac{F(z+\Delta z)-F(z)}{\Delta z} = \frac{1}{\Delta z}\left[\int_{z_0}^{z+\Delta z} f(z^*)\,dz^* - \int_{z_0}^{z} f(z^*)\,dz^*\right] = \frac{1}{\Delta z}\int_{z}^{z+\Delta z} f(z^*)\,dz^*. \tag{5}$$

We now subtract $f(z)$ from (5) and show that the resulting expression approaches zero as $\Delta z \to 0$. The details are as follows.

We keep z fixed. Then we choose $z + \Delta z$ in D so that the whole segment with endpoints z and $z + \Delta z$ is in D (Fig. 352). This can be done because D is a domain, hence it contains a neighborhood of z. We use this segment as the path of integration in (5). Now we subtract $f(z)$. This is a constant because z is kept fixed. Hence we can write

$$\int_{z}^{z+\Delta z} f(z)\,dz^* = f(z)\int_{z}^{z+\Delta z} dz^* = f(z)\,\Delta z. \qquad \text{Thus} \qquad f(z) = \frac{1}{\Delta z}\int_{z}^{z+\Delta z} f(z)\,dz^*.$$

By this trick and from (5) we get a single integral:

$$\frac{F(z+\Delta z)-F(z)}{\Delta z} - f(z) = \frac{1}{\Delta z}\int_{z}^{z+\Delta z} [f(z^*) - f(z)]\,dz^*.$$

Since $f(z)$ is analytic, it is continuous (see Team Project (24d) in Sec. 13.3). An $\epsilon > 0$ being given, we can thus find a $\delta > 0$ such that $|f(z^*) - f(z)| < \epsilon$ when $|z^* - z| < \delta$. Hence, letting $|\Delta z| < \delta$, we see that the *ML*-inequality (Sec. 14.1) yields

$$\left|\frac{F(z+\Delta z)-F(z)}{\Delta z} - f(z)\right| = \frac{1}{|\Delta z|}\left|\int_{z}^{z+\Delta z} [f(z^*) - f(z)]\,dz^*\right| \leqq \frac{1}{|\Delta z|}\epsilon|\Delta z| = \epsilon.$$

By the definition of limit and derivative, this proves that

$$F'(z) = \lim_{\Delta z \to 0} \frac{F(z+\Delta z)-F(z)}{\Delta z} = f(z).$$

Since z is any point in D, this implies that $F(z)$ is analytic in D and is an indefinite integral or antiderivative of $f(z)$ in D, written

$$F(z) = \int f(z)\,dz.$$

Fig. 352. Path of integration

Also, if $G'(z) = f(z)$, then $F'(z) - G'(z) \equiv 0$ in D; hence $F(z) - G(z)$ is constant in D (see Team Project 30 in Problem Set 13.4). That is, two indefinite integrals of $f(z)$ can differ only by a constant. The latter drops out in (9) of Sec. 14.1, so that we can use any indefinite integral of $f(z)$. This proves Theorem 3. ■

Cauchy's Integral Theorem for Multiply Connected Domains

Cauchy's theorem applies to multiply connected domains. We first explain this for a **doubly connected domain** D with outer boundary curve C_1 and inner C_2 (Fig. 353). If a function $f(z)$ is analytic in any domain D^* that contains D and its boundary curves, we claim that

$$\oint_{C_1} f(z)\,dz = \oint_{C_2} f(z)\,dz \qquad \text{(6)} \qquad \text{(Fig. 353)}$$

both integrals being taken counterclockwise (or both clockwise, and regardless of whether or not the full interior of C_2 belongs to D^*).

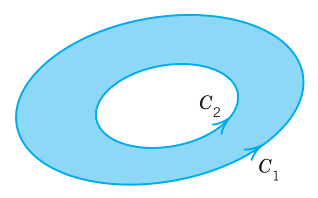


Fig. 353. Paths in (5)

PROOF By two cuts $\tilde{C}_1$ and $\tilde{C}_2$ (Fig. 354) we cut D into two simply connected domains D_1 and D_2 in which and on whose boundaries $f(z)$ is analytic. By Cauchy's integral theorem the integral over the entire boundary of D_1 (taken in the sense of the arrows in Fig. 354) is zero, and so is the integral over the boundary of D_2, and thus their sum. In this sum the integrals over the cuts $\tilde{C}_1$ and $\tilde{C}_2$ cancel because we integrate over them in both directions—this is the key—and we are left with the integrals over C_1 (counterclockwise) and C_2 (clockwise; see Fig. 354); hence by reversing the integration over C_2 (to counterclockwise) we have

$$\oint_{C_1} f\,dz - \oint_{C_2} f\,dz = 0$$

and (6) follows. ■

For domains of higher connectivity the idea remains the same. Thus, for a **triply connected domain** we use three cuts $\tilde{C}_1, \tilde{C}_2, \tilde{C}_3$ (Fig. 355). Adding integrals as before, the integrals over the cuts cancel and the sum of the integrals over C_1 (counterclockwise) and C_2, C_3 (clockwise) is zero. Hence the integral over C_1 equals the sum of the integrals over C_2 and C_3, all three now taken counterclockwise. Similarly for quadruply connected domains, and so on.

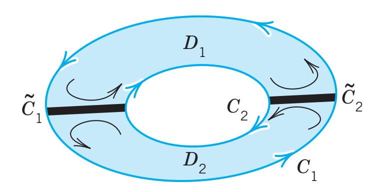


Fig. 354. Doubly connected domain

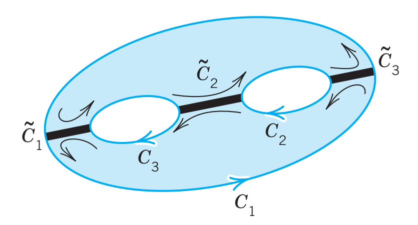


Fig. 355. Triply connected domain

PROBLEM SET 14.2

1–8 COMMENTS ON TEXT AND EXAMPLES

1. Cauchy's Integral Theorem. Verify Theorem 1 for the integral of z^2 over the boundary of the square with vertices $\pm 1 \pm i$. *Hint.* Use deformation.

2. For what contours C will it follow from Theorem 1 that

$$\text{(a)} \quad \int_C \frac{dz}{z} = 0, \qquad \text{(b)} \quad \int_C \frac{\exp(1/z^2)}{z^2 + 16}\, dz = 0\,?$$

3. Deformation principle. Can we conclude from Example 4 that the integral is also zero over the contour in Prob. 1?

4. If the integral of a function over the unit circle equals 2 and over the circle of radius 3 equals 6, can the function be analytic everywhere in the annulus $1 < |z| < 3$?

5. Connectedness. What is the connectedness of the domain in which $(\cos z^2)/(z^4 + 1)$ is analytic?

6. Path independence. Verify Theorem 2 for the integral of e^z from 0 to $1 + i$ **(a)** over the shortest path and **(b)** over the x-axis to 1 and then straight up to $1 + i$.

7. Deformation. Can we conclude in Example 2 that the integral of $1/(z^2 + 4)$ over **(a)** $|z - 2| = 2$ and **(b)** $|z - 2| = 3$ is zero?

8. TEAM EXPERIMENT. Cauchy's Integral Theorem.

(a) Main Aspects. Each of the problems in Examples 1–5 explains a basic fact in connection with Cauchy's theorem. Find five examples of your own, more complicated ones if possible, each illustrating one of those facts.

(b) Partial fractions. Write $f(z)$ in terms of partial fractions and integrate it counterclockwise over the unit circle, where

$$\text{(i)} \quad f(z) = \frac{2z + 3i}{z^2 + \frac{1}{4}} \qquad \text{(ii)} \quad f(z) = \frac{z + 1}{z^2 + 2z}.$$

(c) Deformation of path. Review (c) and (d) of Team Project 34, Sec. 14.1, in the light of the principle of deformation of path. Then consider another family of paths with common endpoints, say, $z(t) = t + ia(t - t^2)$, $0 \leqq t \leqq 1$, a a real constant, and experiment with the integration of analytic and nonanalytic functions of your choice over these paths (e.g., z, $\text{Im}\, z$, z^2, $\text{Re}\, z^2$, $\text{Im}\, z^2$, etc.).

9–19 CAUCHY'S THEOREM APPLICABLE?

Integrate $f(z)$ counterclockwise around the unit circle. Indicate whether Cauchy's integral theorem applies. Show the details.

9. $f(z) = \exp(-z^2)$

10. $f(z) = \tan \frac{1}{4} z$

11. $f(z) = 1/(2z - 1)$

12. $f(z) = \bar{z}^3$

13. $f(z) = 1/(z^4 - 1.1)$

14. $f(z) = 1/\bar{z}$

15. $f(z) = \text{Im}\, z$

16. $f(z) = 1/(\pi z - 1)$

17. $f(z) = 1/|z|^2$

18. $f(z) = 1/(4z - 3)$

19. $f(z) = z^3 \cot z$

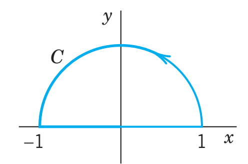
20–30 FURTHER CONTOUR INTEGRALS

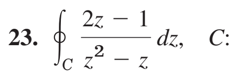
Evaluate the integral. Does Cauchy's theorem apply? Show details.

20. $\oint_C \text{Ln}\,(1 - z)\, dz$, C the boundary of the parallelogram with vertices $\pm i$, $\pm(1 + i)$.

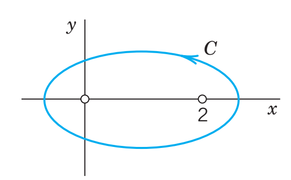
21. $\oint_C \dfrac{dz}{z - 3i}$, C the circle $|z| = \pi$ counterclockwise.

22. $\oint_C \text{Re}\, z\, dz$, C:





23. $\oint_C \dfrac{2z - 1}{z^2 - z}\, dz$, C:



Use partial fractions.

24. $\oint_C \frac{dz}{z^2 - 1}$, C:

Use partial fractions.

25. $\oint_C \frac{e^z}{z}\, dz$, C consists of $|z| = 2$ counterclockwise and $|z| = 1$ clockwise.

26. $\oint_C \coth \frac{1}{2}z\, dz$, C the circle $|z - \frac{1}{2}\pi i| = 1$ clockwise.

27. $\oint_C \frac{\cos z}{z}\, dz$, C consists of $|z| = 1$ counterclockwise and $|z| = 3$ clockwise.

28. $\oint_C \frac{\tan \frac{1}{2}z}{z^4 - 16}\, dz$, C the boundary of the square with vertices ± 1, $\pm i$ clockwise.

29. $\oint_C \frac{\sin z}{z + 2iz}\, dz$, C: $|z - 4 - 2i| = 5.5$ clockwise.

30. $\oint_C \frac{2z^3 + z^2 + 4}{z^4 + 4z^2}\, dz$, C: $|z - 2| = 4$ clockwise. Use partial fractions.

14.3 Cauchy's Integral Formula

Cauchy's integral theorem leads to Cauchy's integral formula. This formula is useful for evaluating integrals as shown in this section. It has other important roles, such as in proving the surprising fact that analytic functions have derivatives of all orders, as shown in the next section, and in showing that all analytic functions have a Taylor series representation (to be seen in Sec. 15.4).

THEOREM 1

Cauchy's Integral Formula

Let $f(z)$ be analytic in a simply connected domain D. Then for any point z_0 in D and any simple closed path C in D that encloses z_0 (Fig. 356),

(1) $$\oint_C \frac{f(z)}{z - z_0}\, dz = 2\pi i f(z_0)$$ **(Cauchy's integral formula)**

the integration being taken counterclockwise. Alternatively (for representing $f(z_0)$ by a contour integral, divide (1) by $2\pi i$),

(1*) $$f(z_0) = \frac{1}{2\pi i}\oint_C \frac{f(z)}{z - z_0}\, dz$$ **(Cauchy's integral formula).**

PROOF By addition and subtraction, $f(z) = f(z_0) + [f(z) - f(z_0)]$. Inserting this into (1) on the left and taking the constant factor $f(z_0)$ out from under the integral sign, we have

(2) $$\oint_C \frac{f(z)}{z - z_0}\, dz = f(z_0)\oint_C \frac{dz}{z - z_0} + \oint_C \frac{f(z) - f(z_0)}{z - z_0}\, dz.$$

The first term on the right equals $f(z_0) \cdot 2\pi i$, which follows from Example 6 in Sec. 14.2 with $m = -1$. If we can show that the second integral on the right is zero, then it would prove the theorem. Indeed, we can. The integrand of the second integral is analytic, except

at z_0. Hence, by (6) in Sec. 14.2, we can replace C by a small circle K of radius ρ and center z_0 (Fig. 357), without altering the value of the integral. Since $f(z)$ is analytic, it is continuous (Team Project 24, Sec. 13.3). Hence, an $\epsilon > 0$ being given, we can find a $\delta > 0$ such that $|f(z) - f(z_0)| < \epsilon$ for all z in the disk $|z - z_0| < \delta$. Choosing the radius ρ of K smaller than δ, we thus have the inequality

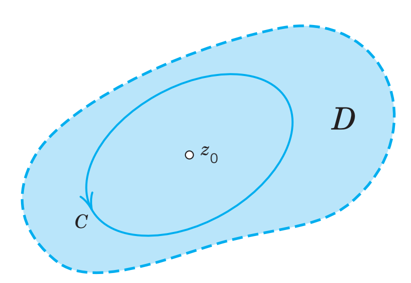


Fig. 356. Cauchy's integral formula

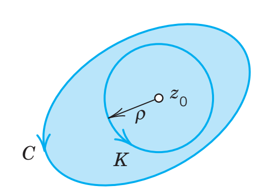


Fig. 357. Proof of Cauchy's integral formula

$$\left|\frac{f(z) - f(z_0)}{z - z_0}\right| < \frac{\epsilon}{\rho}$$

at each point of K. The length of K is $2\pi\rho$. Hence, by the ML-inequality in Sec. 14.1,

$$\left|\oint_K \frac{f(z) - f(z_0)}{z - z_0}\, dz\right| < \frac{\epsilon}{\rho} 2\pi\rho = 2\pi\epsilon.$$

Since ϵ (> 0) can be chosen arbitrarily small, it follows that the last integral in (2) must have the value zero, and the theorem is proved. ■

EXAMPLE 1 Cauchy's Integral Formula

$$\oint_C \frac{e^z}{z - 2}\, dz = 2\pi i e^z \Big|_{z=2} = 2\pi i e^2 = 46.4268i$$

for any contour enclosing $z_0 = 2$ (since e^z is entire), and zero for any contour for which $z_0 = 2$ lies outside (by Cauchy's integral theorem). ■

EXAMPLE 2 Cauchy's Integral Formula

$$\oint_C \frac{z^3 - 6}{2z - i}\, dz = \oint_C \frac{\frac{1}{2}z^3 - 3}{z - \frac{1}{2}i}\, dz$$

$$= 2\pi i \left[\tfrac{1}{2}z^3 - 3\right]\Big|_{z=i/2}$$

$$= \frac{\pi}{8} - 6\pi i \qquad (z_0 = \tfrac{1}{2}i \text{ inside } C). \quad ■$$

EXAMPLE 3 Integration Around Different Contours

Integrate

$$g(z) = \frac{z^2 + 1}{z^2 - 1} = \frac{z^2 + 1}{(z + 1)(z - 1)}$$

counterclockwise around each of the four circles in Fig. 358.

Solution. $g(z)$ is not analytic at -1 and 1. These are the points we have to watch for. We consider each circle separately.

(a) The circle $|z - 1| = 1$ encloses the point $z_0 = 1$ where $g(z)$ is not analytic. Hence in (1) we have to write

$$g(z) = \frac{z^2 + 1}{z^2 - 1} = \frac{z^2 + 1}{z + 1} \frac{1}{z - 1};$$

thus

$$f(z) = \frac{z^2 + 1}{z + 1}$$

and (1) gives

$$\oint_C \frac{z^2 + 1}{z^2 - 1} dz = 2\pi i f(1) = 2\pi i \left[\frac{z^2 + 1}{z + 1}\right]_{z=1} = 2\pi i.$$

(b) gives the same as (a) by the principle of deformation of path.

(c) The function $g(z)$ is as before, but $f(z)$ changes because we must take $z_0 = -1$ (instead of 1). This gives a factor $z - z_0 = z + 1$ in (1). Hence we must write

$$g(z) = \frac{z^2 + 1}{z - 1} \frac{1}{z + 1};$$

thus

$$f(z) = \frac{z^2 + 1}{z - 1}.$$

Compare this for a minute with the previous expression and then go on:

$$\oint_C \frac{z^2 + 1}{z^2 - 1} dz = 2\pi i f(-1) = 2\pi i \left[\frac{z^2 + 1}{z - 1}\right]_{z=-1} = -2\pi i.$$

(d) gives 0. Why? ■

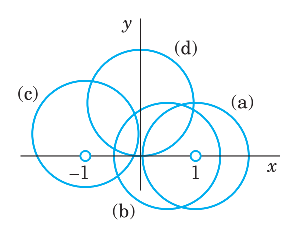


Fig. 358. Example 3

Multiply connected domains can be handled as in Sec. 14.2. For instance, if $f(z)$ is analytic on C_1 and C_2 and in the ring-shaped domain bounded by C_1 and C_2 (Fig. 359) and z_0 is any point in that domain, then

$$(3) \qquad f(z_0) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z)}{z - z_0} dz + \frac{1}{2\pi i} \oint_{C_2} \frac{f(z)}{z - z_0} dz,$$

where the outer integral (over C_1) is taken counterclockwise and the inner clockwise, as indicated in Fig. 359.

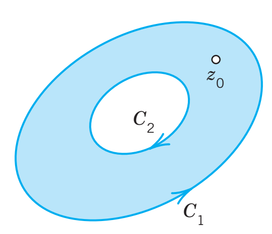


Fig. 359. Formula (3)

PROBLEM SET 14.3

1–4 CONTOUR INTEGRATION

Integrate $z^2/(z^2 - 1)$ by Cauchy's formula counterclockwise around the circle.

1. $|z + 1| = 1$

2. $|z - 1 - i| = \pi/2$

3. $|z + i| = 1.4$

4. $|z + 5 - 5i| = 7$

5–8 Integrate the given function around the unit circle.

5. $(\cos 3z)/(6z)$

6. $e^{2z}/(\pi z - i)$

7. $z^3/(2z - i)$

8. $(z^2 \sin z)/(4z - 1)$

9. CAS EXPERIMENT. Experiment to find out to what extent your CAS can do contour integration. For this, use **(a)** the second method in Sec. 14.1 and **(b)** Cauchy's integral formula.

10. TEAM PROJECT. Cauchy's Integral Theorem. Gain additional insight into the proof of Cauchy's integral theorem by producing (2) with a contour enclosing z_0 (as in Fig. 356) and taking the limit as in the text. Choose

(a) $\displaystyle\oint_C \frac{z^3 - 6}{z - \frac{1}{2}i}\,dz$, **(b)** $\displaystyle\oint_C \frac{\sin z}{z - \frac{1}{2}\pi}\,dz$,

and **(c)** another example of your choice.

11–19 FURTHER CONTOUR INTEGRALS

Integrate counterclockwise or as indicated. Show the details.

11. $\displaystyle\oint_C \frac{dz}{z^2 + 4}$, C: $4x^2 + (y - 2)^2 = 4$

12. $\displaystyle\oint_C \frac{z}{z^2 + 4z + 3}\,dz$, C the circle with center -1 and radius 2

13. $\displaystyle\oint_C \frac{z + 2}{z - 2}\,dz$, C: $|z - 1| = 2$

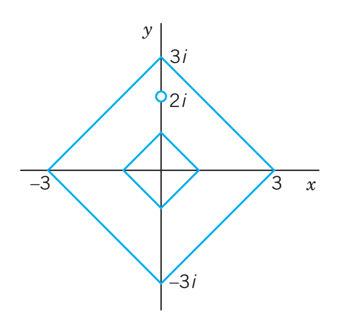
14. $\displaystyle\oint_C \frac{e^z}{ze^z - 2iz}\,dz$, C: $|z| = 0.6$

15. $\displaystyle\oint_C \frac{\cosh (z^2 - \pi i)}{z - \pi i}\,dz$, C the boundary of the square with vertices ± 2, ± 2, $\pm 4i$.

16. $\displaystyle\oint_C \frac{\tan z}{z - i}\,dz$, C the boundary of the triangle with vertices 0 and $\pm 1 + 2i$.

17. $\displaystyle\oint_C \frac{\mathrm{Ln}\,(z + 1)}{z^2 + 1}\,dz$, C: $|z - i| = 1.4$

18. $\displaystyle\oint_C \frac{\sin z}{4z^2 - 8iz}\,dz$, C consists of the boundaries of the squares with vertices ± 3, $\pm 3i$ counterclockwise and ± 1, $\pm i$ clockwise (see figure).



Problem 18

19. $\displaystyle\oint_C \frac{\exp z^2}{z^2(z - 1 - i)}\,dz$, C consists of $|z| = 2$ counterclockwise and $|z| = 1$ clockwise.

20. Show that $\displaystyle\oint_C (z - z_1)^{-1}(z - z_2)^{-1}\,dz = 0$ for a simple closed path C enclosing z_1 and z_2, which are arbitrary.

14.4 Derivatives of Analytic Functions

As mentioned, a surprising fact is that complex analytic functions have derivatives of all orders. This differs completely from real calculus. Even if a real function is once differentiable we cannot conclude that it is twice differentiable nor that any of its higher derivatives exist. This makes the behavior of complex analytic functions simpler than real functions in this aspect. To prove the surprising fact we use Cauchy's integral formula.

THEOREM 1

Derivatives of an Analytic Function

If $f(z)$ is analytic in a domain D, then it has derivatives of all orders in D, which are then also analytic functions in D. The values of these derivatives at a point z_0 in D are given by the formulas

$$f'(z_0) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^2}\, dz \tag{1'}$$

$$f''(z_0) = \frac{2!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^3}\, dz \tag{1''}$$

and in general

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{n+1}}\, dz \qquad (n = 1, 2, \cdots); \tag{1}$$

here C is any simple closed path in D that encloses z_0 and whose full interior belongs to D; and we integrate counterclockwise around C (Fig. 360).

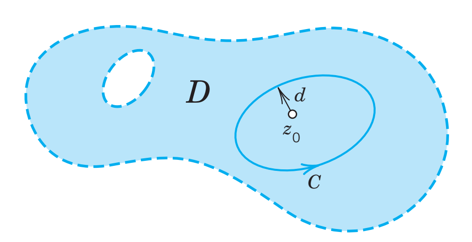


Fig. 360. Theorem 1 and its proof

COMMENT. For memorizing (1), it is useful to observe that these formulas are obtained formally by differentiating the Cauchy formula (1*), Sec. 14.3, under the integral sign *with respect to z_0*.

PROOF We prove (1′), starting from the definition of the derivative

$$f'(z_0) = \lim_{\Delta z \to 0} \frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z}.$$

On the right we represent $f(z_0 + \Delta z)$ and $f(z_0)$ by Cauchy's integral formula:

$$\frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} = \frac{1}{2\pi i \Delta z}\left[\oint_C \frac{f(z)}{z - (z_0 + \Delta z)}\,dz - \oint_C \frac{f(z)}{z - z_0}\,dz\right].$$

We now write the two integrals as a single integral. Taking the common denominator gives the numerator $f(z)\{z - z_0 - [z - (z_0 + \Delta z)]\} = f(z)\,\Delta z$, so that a factor Δz drops out and we get

$$\frac{f(z_0 + \Delta z) - f(z_0)}{\Delta z} = \frac{1}{2\pi i}\oint_C \frac{f(z)}{(z - z_0 - \Delta z)(z - z_0)}\,dz.$$

Clearly, we can now establish (1′) by showing that, as $\Delta z \to 0$, the integral on the right approaches the integral in (1′). To do this, we consider the difference between these two integrals. We can write this difference as a single integral by taking the common denominator and simplifying the numerator (as just before). This gives

$$\oint_C \frac{f(z)}{(z - z_0 - \Delta z)(z - z_0)}\,dz - \oint_C \frac{f(z)}{(z - z_0)^2}\,dz = \oint_C \frac{f(z)\,\Delta z}{(z - z_0 - \Delta z)(z - z_0)^2}\,dz.$$

We show by the *ML*-inequality (Sec. 14.1) that the integral on the right approaches zero as $\Delta z \to 0$.

Being analytic, the function $f(z)$ is continuous on C, hence bounded in absolute value, say, $|f(z)| \leqq K$. Let d be the smallest distance from z_0 to the points of C (see Fig. 360). Then for all z on C,

$$|z - z_0|^2 \geqq d^2, \qquad \text{hence} \qquad \frac{1}{|z - z_0|^2} \leqq \frac{1}{d^2}.$$

Furthermore, by the triangle inequality for all z on C we then also have

$$d \leqq |z - z_0| = |z - z_0 - \Delta z + \Delta z| \leqq |z - z_0 - \Delta z| + |\Delta z|.$$

We now subtract $|\Delta z|$ on both sides and let $|\Delta z| \leqq d/2$, so that $-|\Delta z| \geqq -d/2$. Then

$$\tfrac{1}{2}d \leqq d - |\Delta z| \leqq |z - z_0 - \Delta z|. \qquad \text{Hence} \qquad \frac{1}{|z - z_0 - \Delta z|} \leqq \frac{2}{d}.$$

Let L be the length of C. If $|\Delta z| \leqq d/2$, then by the *ML*-inequality

$$\left|\oint_C \frac{f(z)\,\Delta z}{(z - z_0 - \Delta z)(z - z_0)^2}\,dz\right| \leqq KL\,|\Delta z|\,\frac{2}{d}\cdot\frac{1}{d^2}.$$

This approaches zero as $\Delta z \to 0$. Formula (1′) is proved.

Note that we used Cauchy's integral formula (1*), Sec. 14.3, but if all we had known about $f(z_0)$ is the fact that it can be represented by (1*), Sec. 14.3, our argument would have established the existence of the derivative $f'(z_0)$ of $f(z)$. This is essential to the

continuation and completion of this proof, because it implies that (1″) can be proved by a similar argument, with f replaced by f', and that the general formula (1) follows by induction. ■

Applications of Theorem 1

EXAMPLE 1 **Evaluation of Line Integrals**

From (1′), for any contour enclosing the point πi (counterclockwise)

$$\oint_C \frac{\cos z}{(z-\pi i)^2}\,dz = 2\pi i(\cos z)'\Big|_{z=\pi i} = -2\pi i \sin \pi i = 2\pi \sinh \pi.$$ ■

EXAMPLE 2 From (1″), for any contour enclosing the point $-i$ we obtain by counterclockwise integration

$$\oint_C \frac{z^4-3z^2+6}{(z+i)^3}\,dz = \pi i(z^4-3z^2+6)''\Big|_{z=-i} = \pi i[12z^2-6]_{z=-i} = -18\pi i.$$ ■

EXAMPLE 3 By (1′), for any contour for which 1 lies inside and $\pm 2i$ lie outside (counterclockwise),

$$\oint_C \frac{e^z}{(z-1)^2(z^2+4)}\,dz = 2\pi i\left(\frac{e^z}{z^2+4}\right)'\Bigg|_{z=1}$$

$$= 2\pi i \frac{e^z(z^2+4)-e^z 2z}{(z^2+4)^2}\Bigg|_{z=1} = \frac{6e\pi}{25} i \approx 2.050i.$$ ■

Cauchy's Inequality. Liouville's and Morera's Theorems

We develop other general results about analytic functions, further showing the versatility of Cauchy's integral theorem.

Cauchy's Inequality. Theorem 1 yields a basic inequality that has many applications. To get it, all we have to do is to choose for C in (1) a circle of radius r and center z_0 and apply the ML-inequality (Sec. 14.1); with $|f(z)| \leqq M$ on C we obtain from (1)

$$|f^{(n)}(z_0)| = \frac{n!}{2\pi}\left|\oint_C \frac{f(z)}{(z-z_0)^{n+1}}\,dz\right| \leqq \frac{n!}{2\pi} M \frac{1}{r^{n+1}} 2\pi r.$$

This gives **Cauchy's inequality**

(2)
$$|f^{(n)}(z_0)| \leqq \frac{n!M}{r^n}.$$

To gain a first impression of the importance of this inequality, let us prove a famous theorem on entire functions (definition in Sec. 13.5). (For Liouville, see Sec. 11.5.)

THEOREM 2 **Liouville's Theorem**

If an entire function is bounded in absolute value in the whole complex plane, then this function must be a constant.

PROOF By assumption, $|f(z)|$ is bounded, say, $|f(z)| < K$ for all z. Using (2), we see that $|f'(z_0)| < K/r$. Since $f(z)$ is entire, this holds for every r, so that we can take r as large as we please and conclude that $f'(z_0) = 0$. Since z_0 is arbitrary, $f'(z) = u_x + iv_x = 0$ for all z (see (4) in Sec. 13.4), hence $u_x = v_x = 0$, and $u_y = v_y = 0$ by the Cauchy–Riemann equations. Thus $u = \text{const}$, $v = \text{const}$, and $f = u + iv = \text{const}$ for all z. This completes the proof. ■

Another very interesting consequence of Theorem 1 is

THEOREM 3

Morera's[2] Theorem (Converse of Cauchy's Integral Theorem)

If $f(z)$ is continuous in a simply connected domain D and if

$$\oint_C f(z)\, dz = 0 \tag{3}$$

for every closed path in D, then $f(z)$ is analytic in D.

PROOF In Sec. 14.2 we showed that if $f(z)$ is analytic in a simply connected domain D, then

$$F(z) = \int_{z_0}^{z} f(z^*)\, dz^*$$

is analytic in D and $F'(z) = f(z)$. In the proof we used only the continuity of $f(z)$ and the property that its integral around every closed path in D is zero; from these assumptions we concluded that $F(z)$ is analytic. By Theorem 1, the derivative of $F(z)$ is analytic, that is, $f(z)$ is analytic in D, and Morera's theorem is proved. ■

This completes Chapter 14.

PROBLEM SET 14.4

1–7 CONTOUR INTEGRATION. UNIT CIRCLE

Integrate counterclockwise around the unit circle.

1. $\oint_C \frac{\sin z}{z^4}\, dz$

2. $\oint_C \frac{z^6}{(2z-1)^6}\, dz$

3. $\oint_C \frac{e^z}{z^n}\, dz, \quad n = 1, 2, \cdots$

4. $\oint_C \frac{e^z \cos z}{(z-\pi/4)^3}\, dz$

5. $\oint_C \frac{\cosh 2z}{(z-\frac{1}{2})^4}\, dz$

6. $\oint_C \frac{dz}{(z-2i)^2(z-i/2)^2}$

7. $\oint_C \frac{\cos z}{z^{2n+1}}\, dz, \quad n = 0, 1, \cdots$

8–19 INTEGRATION. DIFFERENT CONTOURS

Integrate. Show the details. *Hint.* Begin by sketching the contour. Why?

8. $\oint_C \frac{z^3 + \sin z}{(z-i)^3}\, dz$, C the boundary of the square with vertices ± 2, $\pm 2i$ counterclockwise.

9. $\oint_C \frac{\tan \pi z}{z^2}\, dz$, C the ellipse $16x^2 + y^2 = 1$ clockwise.

10. $\oint_C \frac{4z^3 - 6}{z(z-1-i)^2}\, dz$, C consists of $|z| = 3$ counterclockwise and $|z| = 1$ clockwise.

[2] GIACINTO MORERA (1856–1909), Italian mathematician who worked in Genoa and Turin.

11. $\oint_C \frac{(1+z)\sin z}{(2z-1)^2}\,dz$, C: $|z-i| = 2$ counterclockwise.

12. $\oint_C \frac{\exp(z^2)}{z(z-2i)^2}\,dz$, C: $z - 3i| = 2$ clockwise.

13. $\oint_C \frac{\operatorname{Ln} z}{(z-2)^2}\,dz$, C: $|z-3| = 2$ counterclockwise.

14. $\oint_C \frac{\operatorname{Ln}(z+3)}{(z-2)(z+1)^2}\,dz$, C the boundary of the square with vertices $\pm 1.5, \pm 1.5i$, counterclockwise.

15. $\oint_C \frac{\cosh 4z}{(z-4)^3}\,dz$, C consists of $|z| = 6$ counterclockwise and $|z-3| = 2$ clockwise.

16. $\oint_C \frac{e^{4z}}{z(z-2i)^2}\,dz$, C consists of $|z-i| = 3$ counterclockwise and $|z| = 1$ clockwise.

17. $\oint_C \frac{e^{-z}\sin z}{(z-4)^3}\,dz$, C consists of $|z| = 5$ counterclockwise and $|z-3| = \frac{3}{2}$ clockwise.

18. $\oint_C \frac{\sinh z}{z^n}\,dz$, C: $|z| = 1$ counterclockwise, n integer.

19. $\oint_C \frac{e^{3z}}{(4z-\pi i)^3}\,dz$, C: $|z| = 1$, counterclockwise.

20. TEAM PROJECT. Theory on Growth

(a) Growth of entire functions. If $f(z)$ is not a constant and is analytic for all (finite) z, and R and M are any positive real numbers (no matter how large), show that there exist values of z for which $|z| > R$ and $|f(z)| > M$. *Hint.* Use Liouville's theorem.

(b) Growth of polynomials. If $f(z)$ is a polynomial of degree $n > 0$ and M is an arbitrary positive real number (no matter how large), show that there exists a positive real number R such that $|f(z)| > M$ for all $|z| > R$.

(c) Exponential function. Show that $f(z) = e^x$ has the property characterized in (a) but does not have that characterized in (b).

(d) Fundamental theorem of algebra. *If $f(z)$ is a polynomial in z, not a constant, then $f(z) = 0$ for at least one value of z.* Prove this. *Hint.* Use (a).

CHAPTER 14 REVIEW QUESTIONS AND PROBLEMS

1. What is a parametric representation of a curve? What is its advantage?

2. What did we assume about paths of integration $z = z(t)$? What is $\dot{z} = dz/dt$ geometrically?

3. State the definition of a complex line integral from memory.

4. Can you remember the relationship between complex and real line integrals discussed in this chapter?

5. How can you evaluate a line integral of an analytic function? Of an arbitrary continous complex function?

6. What value do you get by counterclockwise integration of $1/z$ around the unit circle? You should remember this. It is basic.

7. Which theorem in this chapter do you regard as most important? State it precisely from memory.

8. What is independence of path? Its importance? State a basic theorem on independence of path in complex.

9. What is deformation of path? Give a typical example.

10. Don't confuse Cauchy's integral theorem (also known as ***Cauchy–Goursat theorem***) and Cauchy's integral formula. State both. How are they related?

11. What is a doubly connected domain? How can you extend Cauchy's integral theorem to it?

12. What do you know about derivatives of analytic functions?

13. How did we use integral formulas for derivatives in evaluating integrals?

14. How does the situation for analytic functions differ with respect to derivatives from that in calculus?

15. What is Liouville's theorem? To what complex functions does it apply?

16. What is Morera's theorem?

17. If the integrals of a function $f(z)$ over each of the two boundary circles of an annulus D taken in the same sense have different values, can $f(z)$ be analytic everywhere in D? Give reason.

18. Is $\operatorname{Im} \oint_C f(z)\,dz = \oint_C \operatorname{Im} f(z)\,dz$? Give reason.

19. Is $\left|\oint_C f(z)\,dz\right| = \oint_C |f(z)|\,dz$?

20. How would you find a bound for the left side in Prob. 19?

21–30 INTEGRATION

Integrate by a suitable method.

21. $\int_C z\sinh(z^2)\,dz$ from 0 to $\pi i/2$.

22. $\int_C (|z| + z)\, dz$ clockwise around the unit circle.

23. $\int_C z^{-5} e^z\, dz$ counterclockwise around $|z| = \pi$.

24. $\int_C \text{Re } z\, dz$ from 0 to $3 + 27i$ along $y = x^3$.

25. $\int_C \frac{\tan \pi z}{(z-1)^2}\, dz$ clockwise around $|z - 1| = 0.1$.

26. $\int_C (z^2 + \bar{z}^2)\, dz$ from $z = 0$ horizontally to $z = 2$, then vertically upward to $2 + 2i$.

27. $\int_C (z^2 + \bar{z}^2)\, dz$ from 0 to $2 + 2i$, shortest path.

28. $\oint_C \frac{\text{Ln } z}{(z - 2i)^2}\, dz$ counterclockwise around $|z - 1| = \frac{1}{2}$.

29. $\oint_C \left(\frac{2}{z + 2i} + \frac{1}{z + 4i}\right) dz$ clockwise around $|z - 1| = 2.5$.

30. $\int_C \sin z\, dz$ from 0 to $(1 + i)$.

SUMMARY OF CHAPTER 14
Complex Integration

The **complex line integral** of a function $f(z)$ taken over a path C is denoted by

$$(1) \qquad \int_C f(z)\, dz \quad \text{or, if } C \text{ is closed, also by} \quad \oint_C f(z) \qquad \text{(Sec. 14.1).}$$

If $f(z)$ is analytic in a simply connected domain D, then we can evaluate (1) as in calculus by indefinite integration and substitution of limits, that is,

$$(2) \qquad \int_C f(z)\, dz = F(z_1) - F(z_0) \qquad [F'(z) = f(z)]$$

for every path C in D from a point z_0 to a point z_1 (see Sec. 14.1). These assumptions imply **independence of path**, that is, (2) depends only on z_0 and z_1 (and on $f(z)$, of course) but not on the choice of C (Sec. 14.2). The existence of an $F(z)$ such that $F'(z) = f(z)$ is proved in Sec. 14.2 by Cauchy's integral theorem (see below).

A general method of integration, not restricted to analytic functions, uses the equation $z = z(t)$ of C, where $a \leqq t \leqq b$,

$$(3) \qquad \int_C f(z)\, dz = \int_a^b f(z(t))\dot{z}(t)\, dt \qquad \left(\dot{z} = \frac{dz}{dt}\right).$$

Cauchy's integral theorem is the most important theorem in this chapter. It states that if $f(z)$ is analytic in a simply connected domain D, then for every closed path C in D (Sec. 14.2),

$$(4) \qquad \oint_C f(z)\, dz = 0.$$

Under the same assumptions and for any z_0 in D and closed path C in D containing z_0 in its interior we also have **Cauchy's integral formula**

$$f(z_0) = \frac{1}{2\pi i} \oint_C \frac{f(z)}{z - z_0} \, dz. \tag{5}$$

Furthermore, under these assumptions $f(z)$ has derivatives of all orders in D that are themselves analytic functions in D and (Sec. 14.4)

$$f^{(n)}(z_0) = \frac{n!}{2\pi i} \oint_C \frac{f(z)}{(z - z_0)^{n+1}} \, dz \qquad (n = 1, 2, \cdots). \tag{6}$$

This implies *Morera's theorem* (the converse of Cauchy's integral theorem) and *Cauchy's inequality* (Sec. 14.4), which in turn implies *Liouville's theorem* that an entire function that is bounded in the whole complex plane must be constant.

CHAPTER 15

Power Series, Taylor Series

In Chapter 14, we evaluated complex integrals directly by using Cauchy's integral formula, which was derived from the famous Cauchy integral theorem. We now shift from the approach of Cauchy and Goursat to another approach of evaluating complex integrals, that is, evaluating them by residue integration. This approach, discussed in Chapter 16, first requires a thorough understanding of power series and, in particular, Taylor series. (To develop the theory of residue integration, we still use Cauchy's integral theorem!)

In this chapter, we focus on complex power series and in particular Taylor series. They are analogs of real power series and Taylor series in calculus. Section 15.1 discusses convergence tests for complex series, which are quite similar to those for real series. Thus, if you are familiar with convergence tests from calculus, you may use Sec. 15.1 as a reference section. The main results of this chapter are that complex power series represent analytic functions, as shown in Sec. 15.3, and that, conversely, every analytic function can be represented by power series, called a Taylor series, as shown in Sec. 15.4. The last section (15.5) on uniform convergence is *optional.*

Prerequisite: Chaps. 13, 14.
Sections that may be omitted in a shorter course: 15.1, 15.5.
References and Answers to Problems: App. 1 Part D, App. 2.

15.1 Sequences, Series, Convergence Tests

The basic concepts for ***complex*** sequences and series and tests for convergence and divergence are very similar to those concepts in (real) calculus. ***Thus if you feel at home with real sequences and series and want to take for granted that the ratio test also holds in complex, skip this section and go to Section 15.2.***

Sequences

The basic definitions are as in calculus. An *infinite sequence* or, briefly, a **sequence**, is obtained by assigning to each positive integer n a number z_n, called a **term** of the sequence, and is written

$$z_1, z_2, \cdots \qquad \text{or} \qquad \{z_1, z_2, \cdots\} \qquad \text{or briefly} \qquad \{z_n\}.$$

We may also write $z_0, z_1, \cdots$ or $z_2, z_3, \cdots$ or start with some other integer if convenient.

A **real sequence** is one whose terms are real.

Convergence. A **convergent sequence** $z_1, z_2, \cdots$ is one that has a limit c, written

$$\lim_{n\to\infty} z_n = c \qquad \text{or simply} \qquad z_n \to c.$$

By definition of **limit** this means that for every $\epsilon > 0$ we can find an N such that

$$(1) \qquad |z_n - c| < \epsilon \qquad \text{for all } n > N;$$

geometrically, all terms z_n with $n > N$ lie in the open disk of radius ϵ and center c (Fig. 361) and only finitely many terms do not lie in that disk. [For a *real* sequence, (1) gives an open interval of length 2ϵ and real midpoint c on the real line as shown in Fig. 362.]

A **divergent sequence** is one that does not converge.

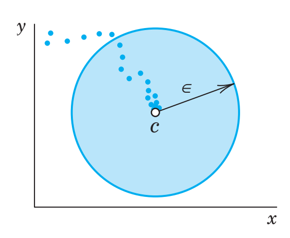


Fig. 361. Convergent complex sequence

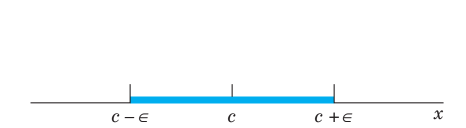


Fig. 362. Convergent real sequence

EXAMPLE 1 **Convergent and Divergent Sequences**

The sequence $\{i^n/n\} = \{i, -\frac{1}{2}, -i/3, \frac{1}{4}, \cdots\}$ is convergent with limit 0.

The sequence $\{i^n\} = \{i, -1, -i, 1, \cdots\}$ is divergent, and so is $\{z_n\}$ with $z_n = (1 + i)^n$. ■

EXAMPLE 2 **Sequences of the Real and the Imaginary Parts**

The sequence $\{z_n\}$ with $z_n = x_n + iy_n = 1 - 1/n^2 + i(2 + 4/n)$ is $6i, \frac{3}{4} + 4i, \frac{8}{9} + 10i/3, \frac{15}{16} + 3i, \cdots$. (Sketch it.) It converges with the limit $c = 1 + 2i$. Observe that $\{x_n\}$ has the limit $1 = \text{Re}\, c$ and $\{y_n\}$ has the limit $2 = \text{Im}\, c$. This is typical. It illustrates the following theorem by which the convergence of a *complex* sequence can be referred back to that of the two *real* sequences of the real parts and the imaginary parts. ■

THEOREM 1 **Sequences of the Real and the Imaginary Parts**

A sequence $z_1, z_2, \cdots, z_n, \cdots$ of complex numbers $z_n = x_n + iy_n$ (where $n = 1, 2, \cdots$) converges to $c = a + ib$ if and only if the sequence of the real parts $x_1, x_2, \cdots$ converges to a and the sequence of the imaginary parts $y_1, y_2, \cdots$ converges to b.

PROOF Convergence $z_n \to c = a + ib$ implies convergence $x_n \to a$ and $y_n \to b$ because if $|z_n - c| < \epsilon$, then z_n lies within the circle of radius ϵ about $c = a + ib$, so that (Fig. 363a)

$$|x_n - a| < \epsilon, \qquad |y_n - b| < \epsilon.$$

Conversely, if $x_n \to a$ and $y_n \to b$ as $n \to \infty$, then for a given $\epsilon > 0$ we can choose N so large that, for every $n > N$,

$$|x_n - a| < \frac{\epsilon}{2}, \qquad |y_n - b| < \frac{\epsilon}{2}.$$

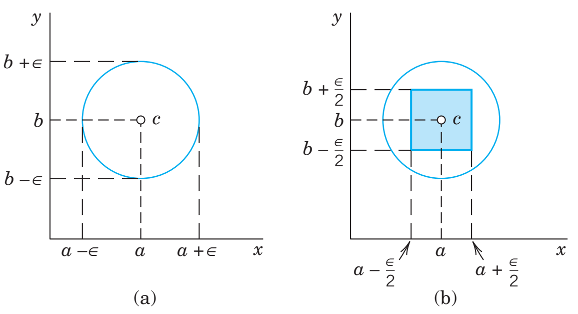


Fig. 363. Proof of Theorem 1

These two inequalities imply that $z_n = x_n + iy_n$ lies in a square with center c and side ϵ. Hence, z_n must lie within a circle of radius ϵ with center c (Fig. 363b). ■

Series

Given a sequence $z_1, z_2, \cdots, z_m, \cdots$, we may form the sequence of the sums

$$s_1 = z_1, \qquad s_2 = z_1 + z_2, \qquad s_3 = z_1 + z_2 + z_3, \quad \cdots$$

and in general

$$s_n = z_1 + z_2 + \cdots + z_n \qquad (n = 1, 2, \cdots). \tag{2}$$

Here s_n is called the ***n*th partial sum** of the *infinite series* or **series**

$$\sum_{m=1}^{\infty} z_m = z_1 + z_2 + \cdots . \tag{3}$$

The $z_1, z_2, \cdots$ are called the **terms** of the series. (Our usual ***summation letter*** is n, unless we need n for another purpose, as here, and we then use m as the summation letter.)

A **convergent series** is one whose sequence of partial sums converges, say,

$$\lim_{n\to\infty} s_n = s. \qquad \text{Then we write} \qquad s = \sum_{m=1}^{\infty} z_m = z_1 + z_2 + \cdots$$

and call s the **sum** or *value* of the series. A series that is not convergent is called a **divergent series**.

If we omit the terms of s_n from (3), there remains

$$R_n = z_{n+1} + z_{n+2} + z_{n+3} + \cdots . \tag{4}$$

This is called the **remainder** *of the series* (3) *after the term* z_n. Clearly, if (3) converges and has the sum s, then

$$s = s_n + R_n, \qquad \text{thus} \qquad R_n = s - s_n.$$

Now $s_n \to s$ by the definition of convergence; hence $R_n \to 0$. In applications, when s is unknown and we compute an approximation s_n of s, then $|R_n|$ is the error, and $R_n \to 0$ means that we can make $|R_n|$ as small as we please, by choosing n large enough.

An application of Theorem 1 to the partial sums immediately relates the convergence of a complex series to that of the two series of its real parts and of its imaginary parts:

THEOREM 2

Real and Imaginary Parts

A series (3) *with* $z_m = x_m + iy_m$ *converges and has the sum* $s = u + iv$ *if and only if* $x_1 + x_2 + \cdots$ *converges and has the sum* u *and* $y_1 + y_2 + \cdots$ *converges and has the sum* v.

Tests for Convergence and Divergence of Series

Convergence tests in complex are practically the same as in calculus. We apply them before we use a series, to make sure that the series converges.

Divergence can often be shown very simply as follows.

THEOREM 3

Divergence

If a series $z_1 + z_2 + \cdots$ *converges, then* $\lim_{m\to\infty} z_m = 0$. *Hence if this does not hold, the series diverges.*

PROOF If $z_1 + z_2 + \cdots$ converges, with the sum s, then, since $z_m = s_m - s_{m-1}$,

$$\lim_{m\to\infty} z_m = \lim_{m\to\infty}(s_m - s_{m-1}) = \lim_{m\to\infty} s_m - \lim_{m\to\infty} s_{m-1} = s - s = 0. \quad \blacksquare$$

CAUTION! $z_m \to 0$ is *necessary* for convergence but *not sufficient*, as we see from the harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \cdots$, which satisfies this condition but diverges, as is shown in calculus (see, for example, Ref. [GenRef11] in App. 1).

The practical difficulty in proving convergence is that, in most cases, the sum of a series is unknown. Cauchy overcame this by showing that a series converges if and only if its partial sums eventually get close to each other:

THEOREM 4

Cauchy's Convergence Principle for Series

A series $z_1 + z_2 + \cdots$ *is convergent if and only if for every given* $\epsilon > 0$ (*no matter how small*) *we can find an* N (*which depends on* ϵ, *in general*) *such that*

$$(5) \qquad |z_{n+1} + z_{n+2} + \cdots + z_{n+p}| < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots$$

The somewhat involved proof is left optional (see App. 4).

Absolute Convergence. A series $z_1 + z_2 + \cdots$ is called **absolutely convergent** if the series of the absolute values of the terms

$$\sum_{m=1}^{\infty} |z_m| = |z_1| + |z_2| + \cdots$$

is convergent.

If $z_1 + z_2 + \cdots$ converges but $|z_1| + |z_2| + \cdots$ diverges, then the series $z_1 + z_2 + \cdots$ is called, more precisely, **conditionally convergent**.

EXAMPLE 3 **A Conditionally Convergent Series**

The series $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + - \cdots$ converges, but only conditionally since the harmonic series diverges, as mentioned above (after Theorem 3). ■

If a series is absolutely convergent, it is convergent.

This follows readily from Cauchy's principle (see Prob. 29). This principle also yields the following general convergence test.

THEOREM 5 **Comparison Test**

If a series $z_1 + z_2 + \cdots$ is given and we can find a convergent series $b_1 + b_2 + \cdots$ with nonnegative real terms such that $|z_1| \leqq b_1, |z_2| \leqq b_2, \cdots$, then the given series converges, even absolutely.

PROOF By Cauchy's principle, since $b_1 + b_2 + \cdots$ converges, for any given $\epsilon > 0$ we can find an N such that

$$b_{n+1} + \cdots + b_{n+p} < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots.$$

From this and $|z_1| \leqq b_1, |z_2| \leqq b_2, \cdots$ we conclude that for those n and p,

$$|z_{n+1}| + \cdots + |z_{n+p}| \leqq b_{n+1} + \cdots + b_{n+p} < \epsilon.$$

Hence, again by Cauchy's principle, $|z_1| + |z_2| + \cdots$ converges, so that $z_1 + z_2 + \cdots$ is absolutely convergent. ■

A good comparison series is the geometric series, which behaves as follows.

THEOREM 6 **Geometric Series**

*The **geometric series***

(6*) $$\sum_{m=0}^{\infty} q^m = 1 + q + q^2 + \cdots$$

converges with the sum $1/(1 - q)$ if $|q| < 1$ and diverges if $|q| \geqq 1$.

PROOF If $|q| \geqq 1$, then $|q^m| \geqq 1$ and Theorem 3 implies divergence.

Now let $|q| < 1$. The nth partial sum is

$$s_n = 1 + q + \cdots + q^n.$$

From this,

$$qs_n = \qquad q + \cdots + q^n + q^{n+1}.$$

On subtraction, most terms on the right cancel in pairs, and we are left with

$$s_n - qs_n = (1 - q)s_n = 1 - q^{n+1}.$$

Now $1 - q \neq 0$ since $q \neq 1$, and we may solve for s_n, finding

$$s_n = \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q} - \frac{q^{n+1}}{1 - q}. \tag{6}$$

Since $|q| < 1$, the last term approaches zero as $n \to \infty$. Hence if $|q| < 1$, the series is convergent and has the sum $1/(1 - q)$. This completes the proof. ■

Ratio Test

This is the most important test in our further work. We get it by taking the geometric series as comparison series $b_1 + b_2 + \cdots$ in Theorem 5:

THEOREM 7

Ratio Test

If a series $z_1 + z_2 + \cdots$ with $z_n \neq 0$ $(n = 1, 2, \cdots)$ has the property that for every n greater than some N,

$$\left|\frac{z_{n+1}}{z_n}\right| \leqq q < 1 \qquad (n > N) \tag{7}$$

(where $q < 1$ is fixed), this series converges absolutely. If for every $n > N$,

$$\left|\frac{z_{n+1}}{z_n}\right| \geqq 1 \qquad (n > N), \tag{8}$$

the series diverges.

PROOF If (8) holds, then $|z_{n+1}| \geqq |z_n|$ for $n > N$, so that divergence of the series follows from Theorem 3.

If (7) holds, then $|z_{n+1}| \leqq |z_n|\, q$ for $n > N$, in particular,

$$|z_{N+2}| \leqq |z_{N+1}|q, \qquad |z_{N+3}| \leqq |z_{N+2}|q \leqq |z_{N+1}|q^2, \qquad \text{etc.,}$$

and in general, $|z_{N+p}| \leqq |z_{N+1}|q^{p-1}$. Since $q < 1$, we obtain from this and Theorem 6

$$|z_{N+1}| + |z_{N+2}| + |z_{N+3}| + \cdots \leqq |z_{N+1}|\,(1 + q + q^2 + \cdots) \leqq |z_{N+1}| \frac{1}{1 - q}.$$

Absolute convergence of $z_1 + z_2 + \cdots$ now follows from Theorem 5. ■

CAUTION! The inequality (7) implies $|z_{n+1}/z_n| < 1$, but this does ***not*** imply convergence, as we see from the harmonic series, which satisfies $z_{n+1}/z_n = n/(n+1) < 1$ for all n but diverges.

If the sequence of the ratios in (7) and (8) converges, we get the more convenient

THEOREM 8

Ratio Test

If a series $z_1 + z_2 + \cdots$ *with* $z_n \neq 0$ $(n = 1, 2, \cdots)$ *is such that* $\lim_{n\to\infty} \left|\frac{z_{n+1}}{z_n}\right| = L$, *then:*

(a) *If* $L < 1$, *the series converges absolutely.*

(b) *If* $L > 1$, *the series diverges.*

(c) *If* $L = 1$, *the series may converge or diverge, so that the test fails and permits no conclusion.*

PROOF **(a)** We write $k_n = |z_{n+1}/z_n|$ and let $L = 1 - b < 1$. Then by the definition of limit, the k_n must eventually get close to $1 - b$, say, $k_n \leqq q = 1 - \frac{1}{2}b < 1$ for all n greater than some N. Convergence of $z_1 + z_2 + \cdots$ now follows from Theorem 7.

(b) Similarly, for $L = 1 + c > 1$ we have $k_n \geqq 1 + \frac{1}{2}c > 1$ for all $n > N^*$ (sufficiently large), which implies divergence of $z_1 + z_2 + \cdots$ by Theorem 7.

(c) The harmonic series $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ has $z_{n+1}/z_n = n/(n+1)$, hence $L = 1$, and diverges. The series

$$1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \frac{1}{25} + \cdots \qquad \text{has} \qquad \frac{z_{n+1}}{z_n} = \frac{n^2}{(n+1)^2},$$

hence also $L = 1$, but it converges. Convergence follows from (Fig. 364)

$$s_n = 1 + \frac{1}{4} + \cdots + \frac{1}{n^2} \leqq 1 + \int_1^n \frac{dx}{x^2} = 2 - \frac{1}{n},$$

so that $s_1, s_2, \cdots$ is a bounded sequence and is monotone increasing (since the terms of the series are all positive); both properties together are sufficient for the convergence of the real sequence $s_1, s_2, \cdots$. (In calculus this is proved by the so-called *integral test*, whose idea we have used.) ■

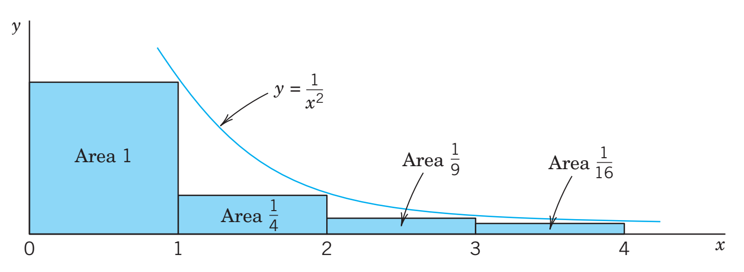


Fig. 364. Convergence of the series $1 + \frac{1}{4} + \frac{1}{9} + \frac{1}{16} + \cdots$

EXAMPLE 4 **Ratio Test**

Is the following series convergent or divergent? (First guess, then calculate.)

$$\sum_{n=0}^{\infty} \frac{(100 + 75i)^n}{n!} = 1 + (100 + 75i) + \frac{1}{2!}(100 + 75i)^2 + \cdots$$

Solution. By Theorem 8, the series is convergent, since

$$\left|\frac{z_{n+1}}{z_n}\right| = \frac{|100 + 75i|^{n+1}/(n + 1)!}{|100 + 75i|^n/n!} = \frac{|100 + 75i|}{n + 1} = \frac{125}{n + 1} \quad \longrightarrow \quad L = 0.$$ ■

EXAMPLE 5 **Theorem 7 More General Than Theorem 8**

Let $a_n = i/2^{3n}$ and $b_n = 1/2^{3n+1}$. Is the following series convergent or divergent?

$$a_0 + b_0 + a_1 + b_1 + \cdots = i + \frac{1}{2} + \frac{i}{8} + \frac{1}{16} + \frac{i}{64} + \frac{1}{128} + \cdots$$

Solution. The ratios of the absolute values of successive terms are $\frac{1}{2}, \frac{1}{4}, \frac{1}{2}, \frac{1}{4}, \cdots$. Hence convergence follows from Theorem 7. Since the sequence of these ratios has no limit, Theorem 8 is not applicable. ■

Root Test

The ratio test and the root test are the two practically most important tests. The ratio test is usually simpler, but the root test is somewhat more general.

THEOREM 9 **Root Test**

If a series $z_1 + z_2 + \cdots$ is such that for every n greater than some N,

$$\sqrt[n]{|z_n|} \leqq q < 1 \qquad (n > N) \tag{9}$$

(where $q < 1$ is fixed), this series converges absolutely. If for infinitely many n,

$$\sqrt[n]{|z_n|} \geqq 1, \tag{10}$$

the series diverges.

PROOF If (9) holds, then $|z_n| \leqq q^n < 1$ for all $n > N$. Hence the series $|z_1| + |z_2| + \cdots$ converges by comparison with the geometric series, so that the series $z_1 + z_2 + \cdots$ converges absolutely. If (10) holds, then $|z_n| \geqq 1$ for infinitely many n. Divergence of $z_1 + z_2 + \cdots$ now follows from Theorem 3. ■

CAUTION! Equation (9) implies $\sqrt[n]{|z_n|} < 1$, but this does not imply convergence, as we see from the harmonic series, which satisfies $\sqrt[n]{1/n} < 1$ (for $n > 1$) but diverges.

If the sequence of the roots in (9) and (10) converges, we more conveniently have

THEOREM 10

Root Test

If a series $z_1 + z_2 + \cdots$ *is such that* $\lim_{n\to\infty} \sqrt[n]{|z_n|} = L$, *then:*

(a) *The series converges absolutely if* $L < 1$.

(b) *The series diverges if* $L > 1$.

(c) *If* $L = 1$, *the test fails; that is, no conclusion is possible.*

PROBLEM SET 15.1

1–10 SEQUENCES

Is the given sequence $z_1, z_2, \cdots, z_n, \cdots$ bounded? Convergent? Find its limit points. Show your work in detail.

1. $z_n = (1 + i)^{2n}/2^n$

2. $z_n = (3 + 4i)^n/n!$

3. $z_n = n\pi/(4 + 2ni)$

4. $z_n = (1 + 2i)^n$

5. $z_n = (-1)^n + 10i$

6. $z_n = (\cos n\pi i)/n$

7. $z_n = n^2 + i/n^2$

8. $z_n = [(1 + 3i)/\sqrt{10}]^n$

9. $z_n = (3 + 3i)^{-n}$

10. $z_n = \sin(\frac{1}{4}n\pi) + i^n$

11. CAS EXPERIMENT. Sequences. Write a program for graphing complex sequences. Use the program to discover sequences that have interesting "geometric" properties, e.g., lying on an ellipse, spiraling to its limit, having infinitely many limit points, etc.

12. Addition of sequences. If $z_1, z_2, \cdots$ converges with the limit l and $z_1^*, z_2^*, \cdots$ converges with the limit l^*, show that $z_1 + z_1^*, z_2 + z_2^*, \cdots$ is convergent with the limit $l + l^*$.

13. Bounded sequence. Show that a complex sequence is bounded if and only if the two corresponding sequences of the real parts and of the imaginary parts are bounded.

14. On Theorem 1. Illustrate Theorem 1 by an example of your own.

15. On Theorem 2. Give another example illustrating Theorem 2.

16–25 SERIES

Is the given series convergent or divergent? Give a reason. Show details.

16. $\sum_{n=0}^{\infty} \frac{(20 + 30i)^n}{n!}$

17. $\sum_{n=2}^{\infty} \frac{(-i)^n}{\ln n}$

18. $\sum_{n=1}^{\infty} n^2\left(\frac{i}{4}\right)^n$

19. $\sum_{n=0}^{\infty} \frac{i^n}{n^2 - i}$

20. $\sum_{n=0}^{\infty} \frac{n + i}{3n^2 + 2i}$

21. $\sum_{n=0}^{\infty} \frac{(\pi + \pi i)^{2n+1}}{(2n + 1)!}$

22. $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}}$

23. $\sum_{n=0}^{\infty} \frac{(-1)^n(1 + i)^{2n}}{(2n)!}$

24. $\sum_{n=1}^{\infty} \frac{(3i)^n n!}{n^n}$

25. $\sum_{n=1}^{\infty} \frac{i^n}{n}$

26. Significance of (7). What is the difference between (7) and just stating $|z_{n+1}/z_n| < 1$?

27. On Theorems 7 and 8. Give another example showing that Theorem 7 is more general than Theorem 8.

28. CAS EXPERIMENT. Series. Write a program for computing and graphing numeric values of the first n partial sums of a series of complex numbers. Use the program to experiment with the rapidity of convergence of series of your choice.

29. Absolute convergence. Show that if a series converges absolutely, it is convergent.

30. Estimate of remainder. Let $|z_{n+1}/z_n| \leqq q < 1$, so that the series $z_1 + z_2 + \cdots$ converges by the ratio test. Show that the remainder $R_n = z_{n+1} + z_{n+2} + \cdots$ satisfies the inequality $|R_n| \leqq |z_{n+1}|/(1 - q)$. Using this, find how many terms suffice for computing the sum s of the series

$$\sum_{n=1}^{\infty} \frac{n + i}{2^n n}$$

with an error not exceeding 0.05 and compute s to this accuracy.

15.2 Power Series

The student should pay close attention to the material because we shall show how power series play an important role in complex analysis. Indeed, they are the most important series in complex analysis because their sums are analytic functions (Theorem 5, Sec. 15.3), and every analytic function can be represented by power series (Theorem 1, Sec. 15.4).

A **power series** *in powers of* $z - z_0$ is a series of the form

$$\sum_{n=0}^{\infty} a_n(z - z_0)^n = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots \tag{1}$$

where z is a complex variable, $a_0, a_1, \cdots$ are complex (or real) constants, called the **coefficients** of the series, and z_0 is a complex (or real) constant, called the **center** of the series. This generalizes real power series of calculus.

If $z_0 = 0$, we obtain as a particular case a *power series in powers of* z:

$$\sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots. \tag{2}$$

Convergence Behavior of Power Series

Power series have variable terms (functions of z), but ***if we fix z, then all the concepts for series with constant terms in the last section apply***. Usually a series with variable terms will converge for some z and diverge for others. For a power series the situation is simple. The series (1) may converge in a disk with center z_0 or in the whole z-plane or only at z_0. We illustrate this with typical examples and then prove it.

EXAMPLE 1 Convergence in a Disk. Geometric Series

The *geometric series*

$$\sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots$$

converges absolutely if $|z| < 1$ and diverges if $|z| \geqq 1$ (see Theorem 6 in Sec. 15.1). ■

EXAMPLE 2 Convergence for Every z

The power series (which will be the Maclaurin series of e^z in Sec. 15.4)

$$\sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \cdots$$

is absolutely convergent for every z. In fact, by the ratio test, for any fixed z,

$$\left| \frac{z^{n+1}/(n+1)!}{z^n/n!} \right| = \frac{|z|}{n+1} \quad \to \quad 0 \qquad \text{as} \qquad n \to \infty. \qquad ■$$

EXAMPLE 3 **Convergence Only at the Center. (Useless Series)**

The following power series converges only at $z = 0$, but diverges for every $z \neq 0$, as we shall show.

$$\sum_{n=0}^{\infty} n!z^n = 1 + z + 2z^2 + 6z^3 + \cdots$$

In fact, from the ratio test we have

$$\left|\frac{(n+1)!z^{n+1}}{n!z^n}\right| = (n+1)|z| \quad \longrightarrow \quad \infty \qquad \text{as} \qquad n \to \infty \qquad (z \text{ fixed and } \neq 0).$$

THEOREM 1 **Convergence of a Power Series**

(a) *Every power series* (1) *converges at the center* z_0.

(b) *If* (1) *converges at a point* $z = z_1 \neq z_0$, *it converges absolutely for every* z *closer to* z_0 *than* z_1, *that is,* $|z - z_0| < |z_1 - z_0|$. See Fig. 365.

(c) *If* (1) *diverges at* $z = z_2$, *it diverges for every* z *farther away from* z_0 *than* z_2. See Fig. 365.

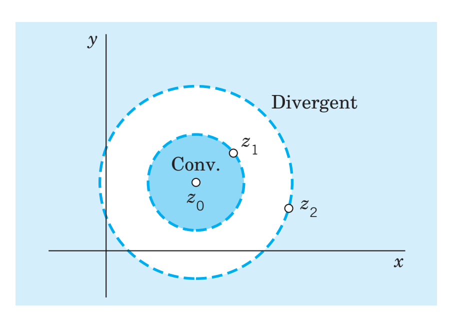


Fig. 365. Theroem 1

PROOF **(a)** For $z = z_0$ the series reduces to the single term a_0.

(b) Convergence at $z = z_1$ gives by Theorem 3 in Sec. 15.1 $a_n(z_1 - z_0)^n \to 0$ as $n \to \infty$. This implies boundedness in absolute value,

$$|a_n(z_1 - z_0)^n| < M \qquad \text{for every } n = 0, 1, \cdots.$$

Multiplying and dividing $a_n(z - z_0)^n$ by $(z_1 - z_0)^n$ we obtain from this

$$|a_n(z - z_0)^n| = \left|a_n(z_1 - z_0)^n \left(\frac{z - z_0}{z_1 - z_0}\right)^n\right| \leqq M \left|\frac{z - z_0}{z_1 - z_0}\right|^n.$$

Summation over n gives

$$\sum_{n=1}^{\infty} |a_n(z - z_0)^n| \leqq M \sum_{n=1}^{\infty} \left|\frac{z - z_0}{z_1 - z_0}\right|^n. \tag{3}$$

Now our assumption $|z - z_0| < |z_1 - z_0|$ implies that $|(z - z_0)/(z_1 - z_0)| < 1$. Hence the series on the right side of (3) is a converging geometric series (see Theorem 6 in

Sec. 15.1). Absolute convergence of (1) as stated in (b) now follows by the comparison test in Sec. 15.1.

(c) If this were false, we would have convergence at a z_3 farther away from z_0 than z_2. This would imply convergence at z_2, by (b), a contradiction to our assumption of divergence at z_2. ■

Radius of Convergence of a Power Series

Convergence for every z (the nicest case, Example 2) or for no $z \neq z_0$ (the useless case, Example 3) needs no further discussion, and we put these cases aside for a moment. We consider the ***smallest*** circle with center z_0 that includes all the points at which a given power series (1) converges. Let R denote its radius. The circle

$$|z - z_0| = R \qquad \text{(Fig. 366)}$$

is called the **circle of convergence** and its radius R the **radius of convergence** of (1). Theorem 1 then implies convergence everywhere within that circle, that is, for all z for which

$$|z - z_0| < R \tag{4}$$

(the open disk with center z_0 and radius R). Also, since R is as *small* as possible, the series (1) diverges for all z for which

$$|z - z_0| > R. \tag{5}$$

No general statements can be made about the convergence of a power series (1) ***on the circle of convergence*** itself. The series (1) may converge at some or all or none of the points. Details will not be important to us. Hence a simple example may just give us the idea.

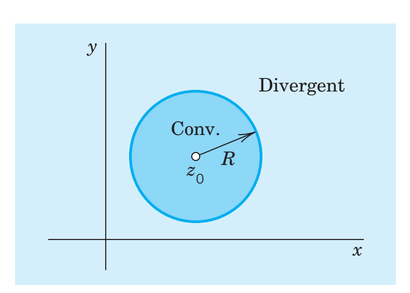


Fig. 366. Circle of convergence

EXAMPLE 4 Behavior on the Circle of Convergence

On the circle of convergence (radius $R = 1$ in all three series),

$\Sigma z^n/n^2$ converges everywhere since $\Sigma 1/n^2$ converges,

$\Sigma z^n/n$ converges at -1 (by Leibniz's test) but diverges at 1,

Σz^n diverges everywhere. ■

Notations $R = \infty$ and $R = 0$. To incorporate these two excluded cases in the present notation, we write

$R = \infty$ if the series (1) converges for all z (as in Example 2),
$R = 0$ if (1) converges only at the center $z = z_0$ (as in Example 3).

These are convenient notations, but nothing else.

Real Power Series. In this case in which powers, coefficients, and center are real, formula (4) gives the **convergence interval** $|x - x_0| < R$ of length $2R$ on the real line.

Determination of the Radius of Convergence from the Coefficients. For this important practical task we can use

THEOREM 2

Radius of Convergence R

Suppose that the sequence $|a_{n+1}/a_n|$, $n = 1, 2, \cdots$, converges with limit L^*. If $L^* = 0$, then $R = \infty$; that is, the power series (1) converges for all z. If $L^* \neq 0$ (hence $L^* > 0$), then

$$R = \frac{1}{L^*} = \lim_{n\to\infty} \left| \frac{a_n}{a_{n+1}} \right| \qquad (6)$$

(Cauchy–Hadamard formula[1]).

If $|a_{n+1}/a_n| \to \infty$, then $R = 0$ (convergence only at the center z_0).

PROOF For (1) the ratio of the terms in the ratio test (Sec. 15.1) is

$$\left| \frac{a_{n+1}(z - z_0)^{n+1}}{a_n(z - z_0)^n} \right| = \left| \frac{a_{n+1}}{a_n} \right| |z - z_0|. \qquad \text{The limit is} \qquad L = L^*|z - z_0|.$$

Let $L^* \neq 0$, thus $L^* > 0$. We have convergence if $L = L^*|z - z_0| < 1$, thus $|z - z_0| < 1/L^*$, and divergence if $|z - z_0| > 1/L^*$. By (4) and (5) this shows that $1/L^*$ is the convergence radius and proves (6).

If $L^* = 0$, then $L = 0$ for every z, which gives convergence for all z by the ratio test. If $|a_{n+1}/a_n| \to \infty$, then $|a_{n+1}/a_n||z - z_0| > 1$ for any $z \neq z_0$ and all sufficiently large n. This implies divergence for all $z \neq z_0$ by the ratio test (Theorem 7, Sec. 15.1). ■

Formula (6) will not help if L^* does not exist, but extensions of Theorem 2 are still possible, as we discuss in Example 6 below.

EXAMPLE 5 **Radius of Convergence**

By (6) the radius of convergence of the power series $\sum_{n=0}^{\infty} \frac{(2n)!}{(n!)^2}(z - 3i)^n$ is

$$R = \lim_{n\to\infty} \left[\frac{(2n!)}{(n!)^2} \bigg/ \frac{(2n+2)!}{((n+1)!)^2} \right] = \lim_{n\to\infty} \left[\frac{(2n!)}{(2n+2)!} \cdot \frac{((n+1)!)^2}{(n!)^2} \right] = \lim_{n\to\infty} \frac{(n+1)^2}{(2n+2)(2n+1)} = \frac{1}{4}.$$

The series converges in the open disk $|z - 3i| < \frac{1}{4}$ of radius $\frac{1}{4}$ and center $3i$. ■

[1] Named after the French mathematicians A. L. CAUCHY (see Sec. 2.5) and JACQUES HADAMARD (1865–1963). Hadamard made basic contributions to the theory of power series and devoted his lifework to partial differential equations.

EXAMPLE 6 **Extension of Theorem 2**

Find the radius of convergence R of the power series

$$\sum_{n=0}^{\infty}\left[1+(-1)^n+\frac{1}{2^n}\right]z^n = 3+\frac{1}{2}z+\left(2+\frac{1}{4}\right)z^2+\frac{1}{8}z^3+\left(2+\frac{1}{16}\right)z^4+\cdots.$$

Solution. The sequence of the ratios $\frac{1}{6}$, $2(2+\frac{1}{4})$, $1/(8(2+\frac{1}{4}))$, $\cdots$ does not converge, so that Theorem 2 is of no help. It can be shown that

$$\text{(6*)}\qquad R = 1/\tilde{L}, \qquad \tilde{L} = \lim_{n\to\infty}\sqrt[n]{|a_n|}.$$

This still does not help here, since $(\sqrt[n]{|a_n|})$ does not converge because $\sqrt[n]{|a_n|} = \sqrt[n]{1/2^n} = \frac{1}{2}$ for odd n, whereas for even n we have

$$\sqrt[n]{|a_n|} = \sqrt[n]{2+1/2^n} \to 1 \qquad \text{as} \qquad n\to\infty,$$

so that $\sqrt[n]{|a_n|}$ has the two limit points $\frac{1}{2}$ and 1. It can further be shown that

$$\text{(6**)}\qquad R = 1/\tilde{l}, \qquad \tilde{l} \text{ the greatest limit point of the sequence } \left\{\sqrt[n]{|a_n|}\right\}.$$

Here $\tilde{l} = I$, so that $R = 1$. *Answer.* The series converges for $|z| < 1$. ■

Summary. Power series converge in an open circular disk or some even for every z (or some only at the center, but they are useless); for the radius of convergence, see (6) or Example 6.

Except for the useless ones, power series have sums that are analytic functions (as we show in the next section); this accounts for their importance in complex analysis.

PROBLEM SET 15.2

1. **Power series.** Are $1/z + z + z^2 + \cdots$ and $z + z^{3/2} + z^2 + z^3 + \cdots$ power series? Explain.
2. **Radius of convergence.** What is it? Its role? What motivates its name? How can you find it?
3. **Convergence.** What are the only basically different possibilities for the convergence of a power series?
4. **On Examples 1–3.** Extend them to power series in powers of $z - 4 + 3\pi i$. Extend Example 1 to the case of radius of convergence 6.
5. **Powers z^{2n}.** Show that if $\Sigma a_n z^n$ has radius of convergence R (assumed finite), then $\Sigma a_n z^{2n}$ has radius of convergence $\sqrt{R}$.

6–18 RADIUS OF CONVERGENCE

Find the center and the radius of convergence.

6. $\sum_{n=0}^{\infty} 4^n(z+1)^n$
7. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!}\left(z-\frac{1}{2}\pi\right)^{2n}$
8. $\sum_{n=0}^{\infty} \frac{n^n}{n!}(z-\pi i)^n$
9. $\sum_{n=0}^{\infty} \frac{n(n-1)}{3^n}(z-i)^{2n}$
10. $\sum_{n=0}^{\infty} \frac{(z-2i)^n}{n^n}$
11. $\sum_{n=0}^{\infty} \left(\frac{2-i}{1+5i}\right)z^n$
12. $\sum_{n=0}^{\infty} \frac{(-1)^n n}{8^n} z^n$
13. $\sum_{n=0}^{\infty} 16^n(z+i)^{4n}$
14. $\sum_{n=0}^{\infty} \frac{(-1)^n}{2^{2n}(n!)^2} z^{2n}$
15. $\sum_{n=0}^{\infty} \frac{(2n)!}{4^n(n!)^2}(z-2i)^n$
16. $\sum_{n=0}^{\infty} \frac{(3n)!}{2^n(n!)^3} z^n$
17. $\sum_{n=1}^{\infty} \frac{2^n}{n(n+1)} z^{2n+1}$
18. $\sum_{n=0}^{\infty} \frac{2(-1)^n}{\sqrt{\pi}(2n+1)n!} z^{2n+1}$
19. **CAS PROJECT. Radius of Convergence.** Write a program for computing R from (6), (6*), or (6**), in

this order, depending on the existence of the limits needed. Test the program on some series of your choice such that all three formulas (6), (6*), and (6**) will come up.

20. TEAM PROJECT. Radius of Convergence.

(a) Understanding (6). Formula (6) for R contains $|a_n/a_{n+1}|$, not $|a_{n+1}/a_n|$. How could you memorize this by using a qualitative argument?

(b) Change of coefficients. What happens to R $(0 < R < \infty)$ if you (i) multiply all a_n by $k \neq 0$, (ii) multiply all a_n by $k^n \neq 0$, (iii) replace a_n by $1/a_n$? Can you think of an application of this?

(c) Understanding Example 6, which extends Theorem 2 to nonconvergent cases of a_n/a_{n+1}. Do you understand the principle of "mixing" by which Example 6 was obtained? Make up further examples.

(d) Understanding (b) and (c) in Theorem 1. Does there exist a power series in powers of z that converges at $z = 30 + 10i$ and diverges at $z = 31 - 6i$? Give reason.

15.3 Functions Given by Power Series

Here, our main goal is to show that power series represent analytic functions. This fact (Theorem 5) and the fact that power series behave nicely under addition, multiplication, differentiation, and integration accounts for their usefulness.

To simplify the formulas in this section, we take $z_0 = 0$ and write

$$\sum_{n=0}^{\infty} a_n z^n. \tag{1}$$

There is no loss of generality because a series in powers of $\hat{z} - z_0$ with any z_0 can always be reduced to the form (1) if we set $\hat{z} - z_0 = z$.

Terminology and Notation. If any given power series (1) has a nonzero radius of convergence R (thus $R > 0$), its sum is a function of z, say $f(z)$. Then we write

$$f(z) = \sum_{n=0}^{\infty} a_n z^n = a_0 + a_1 z + a_2 z^2 + \cdots \qquad (|z| < R). \tag{2}$$

We say that $f(z)$ *is* **represented** *by the power series* or that *it is* **developed** *in the power series.* For instance, the geometric series *represents* the function $f(z) = 1/(1 - z)$ in the interior of the unit circle $|z| = 1$. (See Theorem 6 in Sec. 15.1.)

Uniqueness of a Power Series Representation. This is our next goal. It means that *a function $f(z)$ cannot be represented by two different power series with the same center.* We claim that if $f(z)$ can at all be developed in a power series with center z_0, the development is unique. This important fact is frequently used in complex analysis (as well as in calculus). We shall prove it in Theorem 2. The proof will follow from

THEOREM 1

Continuity of the Sum of a Power Series

If a function $f(z)$ can be represented by a power series (2) *with radius of convergence $R > 0$, then $f(z)$ is continuous at $z = 0$.*

PROOF From (2) with $z = 0$ we have $f(0) = a_0$. Hence by the definition of continuity we must show that $\lim_{z \to 0} f(z) = f(0) = a_0$. That is, we must show that for a given $\epsilon > 0$ there is a $\delta > 0$ such that $|z| < \delta$ implies $|f(z) - a_0| < \epsilon$. Now (2) converges absolutely for $|z| \leqq r$ with any r such that $0 < r < R$, by Theorem 1 in Sec. 15.2. Hence the series

$$\sum_{n=1}^{\infty} |a_n| r^{n-1} = \frac{1}{r} \sum_{n=1}^{\infty} |a_n| r^n$$

converges. Let $S \neq 0$ be its sum. ($S = 0$ is trivial.) Then for $0 < |z| \leqq r$,

$$|f(z) - a_0| = \left| \sum_{n=1}^{\infty} a_n z^n \right| \leqq |z| \sum_{n=1}^{\infty} |a_n| |z|^{n-1} \leqq |z| \sum_{n=1}^{\infty} |a_n| r^{n-1} = |z| S$$

and $|z|S < \epsilon$ when $|z| < \delta$, where $\delta > 0$ is less than r and less than ϵ/S. Hence $|z|S < \delta S < (\epsilon/S)S = \epsilon$. This proves the theorem. ■

From this theorem we can now readily obtain the desired uniqueness theorem (again assuming $z_0 = 0$ without loss of generality):

THEOREM 2

Identity Theorem for Power Series. Uniqueness

Let the power series $a_0 + a_1 z + a_2 z^2 + \cdots$ and $b_0 + b_1 z + b_2 z^2 + \cdots$ both be convergent for $|z| < R$, where R is positive, and let them both have the same sum for all these z. Then the series are identical, that is, $a_0 = b_0, a_1 = b_1, a_2 = b_2, \cdots$.

*Hence if a function $f(z)$ can be represented by a power series with any center z_0, this representation is **unique**.*

PROOF We proceed by induction. By assumption,

$$a_0 + a_1 z + a_2 z^2 + \cdots = b_0 + b_1 z + b_2 z^2 + \cdots \qquad (|z| < R).$$

The sums of these two power series are continuous at $z = 0$, by Theorem 1. Hence if we consider $|z| > 0$ and let $z \to 0$ on both sides, we see that $a_0 = b_0$: the assertion is true for $n = 0$. Now assume that $a_n = b_n$ for $n = 0, 1, \cdots, m$. Then on both sides we may omit the terms that are equal and divide the result by z^{m+1} ($\neq 0$); this gives

$$a_{m+1} + a_{m+2} z + a_{m+3} z^2 + \cdots = b_{m+1} + b_{m+2} z + b_{m+3} z^2 + \cdots.$$

Similarly as before by letting $z \to 0$ we conclude from this that $a_{m+1} = b_{m+1}$. This completes the proof. ■

Operations on Power Series

Interesting in itself, this discussion will serve as a preparation for our main goal, namely, to show that functions represented by power series are analytic.

Termwise addition or subtraction of two power series with radii of convergence R_1 and R_2 yields a power series with radius of convergence at least equal to the smaller of R_1 and R_2. *Proof.* Add (or subtract) the partial sums s_n and s_n^* term by term and use $\lim (s_n \pm s_n^*) = \lim s_n \pm \lim s_n^*$.

Termwise multiplication of two power series

$$f(z) = \sum_{k=0}^{\infty} a_k z^k = a_0 + a_1 z + \cdots$$

and

$$g(z) = \sum_{m=0}^{\infty} b_m z^m = b_0 + b_1 z + \cdots$$

means the multiplication of each term of the first series by each term of the second series and the collection of like powers of z. This gives a power series, which is called the **Cauchy product** of the two series and is given by

$$a_0 b_0 + (a_0 b_1 + a_1 b_0)z + (a_0 b_2 + a_1 b_1 + a_2 b_0)z^2 + \cdots$$
$$= \sum_{n=0}^{\infty} (a_0 b_n + a_1 b_{n-1} + \cdots + a_n b_0) z^n.$$

We mention without proof that this power series converges absolutely for each z within the smaller circle of convergence of the two given series and has the sum $s(z) = f(z)g(z)$. For a proof, see [D5] listed in App. 1.

Termwise differentiation and integration of power series is permissible, as we show next. We call **derived series** *of the power series* (1) the power series obtained from (1) by termwise differentiation, that is,

$$\sum_{n=1}^{\infty} n a_n z^{n-1} = a_1 + 2a_2 z + 3a_3 z^2 + \cdots. \tag{3}$$

THEOREM 3

Termwise Differentiation of a Power Series

The derived series of a power series has the same radius of convergence as the original series.

PROOF This follows from (6) in Sec. 15.2 because

$$\lim_{n\to\infty} \frac{n|a_n|}{(n+1)|a_{n+1}|} = \lim_{n\to\infty} \frac{n}{n+1} \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right| = \lim_{n\to\infty} \left|\frac{a_n}{a_{n+1}}\right|$$

or, if the limit does not exist, from (6**) in Sec. 15.2 by noting that $\sqrt[n]{n} \to 1$ as $n \to \infty$. ■

EXAMPLE 1 Application of Theorem 3

Find the radius of convergence R of the following series by applying Theorem 3.

$$\sum_{n=2}^{\infty} \binom{n}{2} z^n = z^2 + 3z^3 + 6z^4 + 10z^5 + \cdots .$$

Solution. Differentiate the geometric series twice term by term and multiply the result by $z^2/2$. This yields the given series. Hence $R = 1$ by Theorem 3. ■

THEOREM 4 Termwise Integration of Power Series

The power series

$$\sum_{n=0}^{\infty} \frac{a_n}{n+1} z^{n+1} = a_0 z + \frac{a_1}{2} z^2 + \frac{a_2}{3} z^3 + \cdots$$

obtained by integrating the series $a_0 + a_1 z + a_2 z^2 + \cdots$ term by term has the same radius of convergence as the original series.

The proof is similar to that of Theorem 3.

With the help of Theorem 3, we establish the main result in this section.

Power Series Represent Analytic Functions

THEOREM 5 Analytic Functions. Their Derivatives

A power series with a nonzero radius of convergence R represents an analytic function at every point interior to its circle of convergence. The derivatives of this function are obtained by differentiating the original series term by term. All the series thus obtained have the same radius of convergence as the original series. Hence, by the first statement, each of them represents an analytic function.

PROOF **(a)** We consider any power series (1) with positive radius of convergence R. Let $f(z)$ be its sum and $f_1(z)$ the sum of its derived series; thus

$$f(z) = \sum_{n=0}^{\infty} a_n z^n \qquad \text{and} \qquad f_1(z) = \sum_{n=1}^{\infty} n a_n z^{n-1}. \tag{4}$$

We show that $f(z)$ is analytic and has the derivative $f_1(z)$ in the interior of the circle of convergence. We do this by proving that for any fixed z with $|z| < R$ and $\Delta z \to 0$ the difference quotient $[f(z + \Delta z) - f(z)]/\Delta z$ approaches $f_1(z)$. By termwise addition we first have from (4)

$$\frac{f(z+\Delta z) - f(z)}{\Delta z} - f_1(z) = \sum_{n=2}^{\infty} a_n \left[\frac{(z+\Delta z)^n - z^n}{\Delta z} - n z^{n-1} \right]. \tag{5}$$

Note that the summation starts with 2, since the constant term drops out in taking the difference $f(z + \Delta z) - f(z)$, and so does the linear term when we subtract $f_1(z)$ from the difference quotient.

(b) We claim that the series in (5) can be written

$$\sum_{n=2}^{\infty} a_n \, \Delta z[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n - 2)z^{n-3}(z + \Delta z) + (n - 1)z^{n-2}]. \tag{6}$$

The somewhat technical proof of this is given in App. 4.

(c) We consider (6). The brackets contain $n - 1$ terms, and the largest coefficient is $n - 1$. Since $(n - 1)^2 \leqq n(n - 1)$, we see that for $|z| \leqq R_0$ and $|z + \Delta z| \leqq R_0$, $R_0 < R$, the absolute value of this series (6) cannot exceed

$$|\Delta z| \sum_{n=2}^{\infty} |a_n| n(n - 1)R_0^{n-2}. \tag{7}$$

This series with a_n instead of $|a_n|$ is the second derived series of (2) at $z = R_0$ and converges absolutely by Theorem 3 of this section and Theorem 1 of Sec. 15.2. Hence our present series (7) converges. Let the sum of (7) (without the factor $|\Delta z|$) be $K(R_0)$. Since (6) is the right side of (5), our present result is

$$\left| \frac{f(z + \Delta z) - f(z)}{\Delta z} - f_1(z) \right| \leqq |\Delta z| K(R_0).$$

Letting $\Delta z \rightarrow 0$ and noting that R_0 $(< R)$ is arbitrary, we conclude that $f(z)$ is analytic at any point interior to the circle of convergence and its derivative is represented by the derived series. From this the statements about the higher derivatives follow by induction. ■

Summary. The results in this section show that power series are about as nice as we could hope for: we can differentiate and integrate them term by term (Theorems 3 and 4). Theorem 5 accounts for the great importance of power series in complex analysis: the sum of such a series (with a positive radius of convergence) is an analytic function and has derivatives of all orders, which thus in turn are analytic functions. But this is only part of the story. In the next section we show that, conversely, *every* given analytic function $f(z)$ can be represented by power series, called ***Taylor series*** and being the complex analog of the real Taylor series of calculus.

PROBLEM SET 15.3

1. **Relation to Calculus.** Material in this section generalizes calculus. Give details.
2. **Termwise addition.** Write out the details of the proof on termwise addition and subtraction of power series.
3. **On Theorem 3.** Prove that $\sqrt[n]{n} \rightarrow 1$ as $n \rightarrow \infty$, as claimed.
4. **Cauchy product.** Show that $(1 - z)^{-2} = \sum_{n=0}^{\infty} (n + 1)z^n$ **(a)** by using the Cauchy product, **(b)** by differentiating a suitable series.

5–15 RADIUS OF CONVERGENCE BY DIFFERENTIATION OR INTEGRATION

Find the radius of convergence in two ways: **(a)** directly by the Cauchy–Hadamard formula in Sec. 15.2, and **(b)** from a series of simpler terms by using Theorem 3 or Theorem 4.

5. $\displaystyle\sum_{n=2}^{\infty} \frac{n(n - 1)}{2^n} (z - 2i)^n$
6. $\displaystyle\sum_{n=0}^{\infty} \frac{(-1)^n}{2n + 1} \left(\frac{z}{2\pi}\right)^{2n+1}$
7. $\displaystyle\sum_{n=1}^{\infty} \frac{n}{3^n} (z + 2i)^{2n}$
8. $\displaystyle\sum_{n=1}^{\infty} \frac{5^n}{n(n + 1)} z^n$

9. $\sum_{n=1}^{\infty} \frac{(-2)^n}{n(n+1)(n+2)} z^{2n}$

10. $\sum_{n=k}^{\infty} \binom{n}{k} \left(\frac{z}{2}\right)^n$

11. $\sum_{n=1}^{\infty} \frac{3^n n(n+1)}{7^n} (z+2)^{2n}$

12. $\sum_{n=1}^{\infty} \frac{2n(2n-1)}{n^n} z^{2n-2}$

13. $\sum_{n=0}^{\infty} \left[\binom{n+k}{k}\right]^{-1} z^{n+k}$

14. $\sum_{n=0}^{\infty} \binom{n+m}{m} z^n$

15. $\sum_{n=2}^{\infty} \frac{4^n n(n-1)}{3^n} (z-i)^n$

16–20 APPLICATIONS OF THE IDENTITY THEOREM

State clearly and explicitly where and how you are using Theorem 2.

16. Even functions. If $f(z)$ in (2) is *even* (i.e., $f(-z) = f(z)$), show that $a_n = 0$ for odd n. Give examples.

17. Odd function. If $f(z)$ in (2) is *odd* (i.e., $f(-z) = -f(z)$), show that $a_n = 0$ for even n. Give examples.

18. Binomial coefficients. Using $(1+z)^p(1+z)^q = (1+z)^{p+q}$, obtain the basic relation

$$\sum_{n=0}^{r} \binom{p}{n}\binom{q}{r-n} = \binom{p+q}{r}.$$

19. Find applications of Theorem 2 in differential equations and elsewhere.

20. TEAM PROJECT. Fibonacci numbers.[2] **(a)** The Fibonacci numbers are recursively defined by $a_0 = a_1 = 1$, $a_{n+1} = a_n + a_{n-1}$ if $n = 1, 2, \cdots$. Find the limit of the sequence (a_{n+1}/a_n).

(b) Fibonacci's rabbit problem. Compute a list of $a_1, \cdots, a_{12}$. Show that $a_{12} = 233$ is the number of pairs of rabbits after 12 months if initially there is 1 pair and each pair generates 1 pair per month, beginning in the second month of existence (no deaths occurring).

(c) Generating function. Show that the *generating function* of the **Fibonacci numbers** is $f(z) = 1/(1 - z - z^2)$; that is, if a power series (1) represents this $f(z)$, its coefficients must be the Fibonacci numbers and conversely. *Hint.* Start from $f(z)(1 - z - z^2) = 1$ and use Theorem 2.

15.4 Taylor and Maclaurin Series

The **Taylor series**[3] of a function $f(z)$, the complex analog of the real Taylor series is

(1) $$f(z) = \sum_{n=1}^{\infty} a_n(z-z_0)^n \qquad \text{where} \qquad a_n = \frac{1}{n!} f^{(n)}(z_0)$$

or, by (1), Sec. 14.4,

(2) $$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^*-z_0)^{n+1}}\, dz^*.$$

In (2) we integrate counterclockwise around a simple closed path C that contains z_0 in its interior and is such that $f(z)$ is analytic in a domain containing C and every point inside C.

A **Maclaurin series**[3] is a Taylor series with center $z_0 = 0$.

[2] LEONARDO OF PISA, called FIBONACCI (= son of Bonaccio), about 1180–1250, Italian mathematician, credited with the first renaissance of mathematics on Christian soil.

[3] BROOK TAYLOR (1685–1731), English mathematician who introduced real Taylor series. COLIN MACLAURIN (1698–1746), Scots mathematician, professor at Edinburgh.

The **remainder** of the Taylor series (1) after the term $a_n(z - z_0)^n$ is

$$R_n(z) = \frac{(z - z_0)^{n+1}}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}(z^* - z)} dz^* \tag{3}$$

(proof below). Writing out the corresponding partial sum of (1), we thus have

$$\begin{aligned} f(z) = f(z_0) &+ \frac{z - z_0}{1!} f'(z_0) + \frac{(z - z_0)^2}{2!} f''(z_0) + \cdots \\ &+ \frac{(z - z_0)^n}{n!} f^{(n)}(z_0) + R_n(z). \end{aligned} \tag{4}$$

This is called **Taylor's formula** *with remainder.*

We see that ***Taylor series are power series***. From the last section we know that power series represent analytic functions. And we now show that *every* analytic function can be represented by power series, namely, by Taylor series (with various centers). This makes Taylor series very important in complex analysis. Indeed, they are more fundamental in complex analysis than their real counterparts are in calculus.

THEOREM 1

Taylor's Theorem

Let $f(z)$ be analytic in a domain D, and let $z = z_0$ be any point in D. Then there exists precisely one Taylor series (1) with center z_0 that represents $f(z)$. This representation is valid in the largest open disk with center z_0 in which $f(z)$ is analytic. The remainders $R_n(z)$ of (1) can be represented in the form (3). The coefficients satisfy the inequality

$$|a_n| \leqq \frac{M}{r^n} \tag{5}$$

where M is the maximum of $|f(z)|$ on a circle $|z - z_0| = r$ in D whose interior is also in D.

PROOF The key tool is Cauchy's integral formula in Sec. 14.3; writing z and z^* instead of z_0 and z (so that z^* is the variable of integration), we have

$$f(z) = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{z^* - z} dz^*. \tag{6}$$

z lies inside C, for which we take a circle of radius r with center z_0 and interior in D (Fig. 367). We develop $1/(z^* - z)$ in (6) in powers of $z - z_0$. By a ***standard algebraic manipulation*** (worth remembering!) we first have

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0 - (z - z_0)} = \frac{1}{(z^* - z_0)\left(1 - \dfrac{z - z_0}{z^* - z_0}\right)}. \tag{7}$$

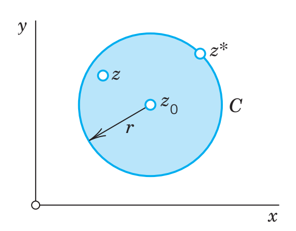


Fig. 367. Cauchy formula (6)

For later use we note that since z^* is on C while z is inside C, we have

$$\left|\frac{z - z_0}{z^* - z_0}\right| < 1. \qquad \text{(Fig. 367).} \tag{7*}$$

To (7) we now apply the sum formula for a finite geometric sum

$$1 + q + \cdots + q^n = \frac{1 - q^{n+1}}{1 - q} = \frac{1}{1 - q} - \frac{q^{n+1}}{1 - q} \qquad (q \neq 1), \tag{8*}$$

which we use in the form (take the last term to the other side and interchange sides)

$$\frac{1}{1 - q} = 1 + q + \cdots + q^n + \frac{q^{n+1}}{1 - q}. \tag{8}$$

Applying this with $q = (z - z_0)/(z^* - z_0)$ to the right side of (7), we get

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0}\left[1 + \frac{z - z_0}{z^* - z_0} + \left(\frac{z - z_0}{z^* - z_0}\right)^2 + \cdots + \left(\frac{z - z_0}{z^* - z_0}\right)^n\right]$$
$$+ \frac{1}{z^* - z}\left(\frac{z - z_0}{z^* - z_0}\right)^{n+1}.$$

We insert this into (6). Powers of $z - z_0$ do not depend on the variable of integration z^*, so that we may take them out from under the integral sign. This yields

$$f(z) = \frac{1}{2\pi i}\oint_C \frac{f(z^*)}{z^* - z_0}\,dz^* + \frac{z - z_0}{2\pi i}\oint_C \frac{f(z^*)}{(z^* - z_0)^2}\,dz^* + \cdots$$
$$\cdots + \frac{(z - z_0)^n}{2\pi i}\oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\,dz^* + R_n(z)$$

with $R_n(z)$ given by (3). The integrals are those in (2) related to the derivatives, so that we have proved the Taylor formula (4).

Since analytic functions have derivatives of all orders, we can take n in (4) as large as we please. If we let n approach infinity, we obtain (1). Clearly, (1) will converge and represent $f(z)$ if and only if

$$\lim_{n \to \infty} R_n(z) = 0. \tag{9}$$

We prove (9) as follows. Since z^* lies on C, whereas z lies inside C (Fig. 367), we have $|z^* - z| > 0$. Since $f(z)$ is analytic inside and on C, it is bounded, and so is the function $f(z^*)/(z^* - z)$, say,

$$\left|\frac{f(z^*)}{z^* - z}\right| \leqq \tilde{M}$$

for all z^* on C. Also, C has the radius $r = |z^* - z_0|$ and the length $2\pi r$. Hence by the ML-inequality (Sec. 14.1) we obtain from (3)

$$\begin{aligned} |R_n| &= \frac{|z - z_0|^{n+1}}{2\pi} \left| \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}(z^* - z)} dz^* \right| \\ &\leqq \frac{|z - z_0|^{n+1}}{2\pi} \tilde{M} \frac{1}{r^{n+1}} 2\pi r = \tilde{M} \left| \frac{z - z_0}{r} \right|^{n+1}. \end{aligned} \tag{10}$$

Now $|z - z_0| < r$ because z lies *inside* C. Thus $|z - z_0|/r < 1$, so that the right side approaches 0 as $n \to \infty$. This proves that the Taylor series converges and has the sum $f(z)$. Uniqueness follows from Theorem 2 in the last section. Finally, (5) follows from a_n in (1) and the Cauchy inequality in Sec. 14.4. This proves Taylor's theorem. ■

Accuracy of Approximation. We can achieve any preassinged accuracy in approximating $f(z)$ by a partial sum of (1) by choosing n large enough. This is the practical use of formula (9).

Singularity, Radius of Convergence. On the circle of convergence of (1) there is at least one **singular point** of $f(z)$, that is, a point $z = c$ at which $f(z)$ is not analytic (but such that every disk with center c contains points at which $f(z)$ *is* analytic). We also say that $f(z)$ **is singular** at c or **has a singularity** at c. Hence the radius of convergence R of (1) is usually equal to the distance from z_0 to the nearest singular point of $f(z)$.

(Sometimes R can be greater than that distance: Ln z is singular on the negative real axis, whose distance from $z_0 = -1 + i$ is 1, but the Taylor series of Ln z with center $z_0 = -1 + i$ has radius of convergence $\sqrt{2}$.)

Power Series as Taylor Series

Taylor series are power series—of course! Conversely, we have

THEOREM 2

Relation to the Previous Section

A power series with a nonzero radius of convergence is the Taylor series of its sum.

PROOF Given the power series

$$f(z) = a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + a_3(z - z_0)^3 + \cdots.$$

Then $f(z_0) = a_0$. By Theorem 5 in Sec. 15.3 we obtain

$$f'(z) = a_1 + 2a_2(z - z_0) + 3a_3(z - z_0)^2 + \cdots, \qquad \text{thus} \qquad f'(z_0) = a_1$$
$$f''(z) = 2a_2 + 3 \cdot 2(z - z_0) + \cdots, \qquad \text{thus} \qquad f''(z_0) = 2!a_2$$

and in general $f^{(n)}(z_0) = n!a_n$. With these coefficients the given series becomes the Taylor series of $f(z)$ with center z_0. ■

Comparison with Real Functions. One surprising property of complex analytic functions is that they have derivatives of all orders, and now we have discovered the other surprising property that they can always be represented by power series of the form (1). This is not true in general for ***real functions;*** there are real functions that have derivatives of all orders but cannot be represented by a power series. (Example: $f(x) = \exp(-1/x^2)$ if $x \neq 0$ and $f(0) = 0$; this function cannot be represented by a Maclaurin series in an open disk with center 0 because all its derivatives at 0 are zero.)

Important Special Taylor Series

These are as in calculus, with x replaced by complex z. Can you see why? (*Answer.* The coefficient formulas are the same.)

EXAMPLE 1 Geometric Series

Let $f(z) = 1/(1 - z)$. Then we have $f^{(n)}(z) = n!/(1 - z)^{n+1}$, $f^{(n)}(0) = n!$. Hence the Maclaurin expansion of $1/(1 - z)$ is the geometric series

(11)
$$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n = 1 + z + z^2 + \cdots \qquad (|z| < 1).$$

$f(z)$ is singular at $z = 1$; this point lies on the circle of convergence. ■

EXAMPLE 2 Exponential Function

We know that the exponential function e^z (Sec. 13.5) is analytic for all z, and $(e^z)' = e^z$. Hence from (1) with $z_0 = 0$ we obtain the Maclaurin series

(12)
$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \cdots.$$

This series is also obtained if we replace x in the familiar Maclaurin series of e^x by z.

Furthermore, by setting $z = iy$ in (12) and separating the series into the real and imaginary parts (see Theorem 2, Sec. 15.1) we obtain

$$e^{iy} = \sum_{n=0}^{\infty} \frac{(iy)^n}{n!} = \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k}}{(2k)!} + i \sum_{k=0}^{\infty} (-1)^k \frac{y^{2k+1}}{(2k+1)!}.$$

Since the series on the right are the familiar Maclaurin series of the real functions $\cos y$ and $\sin y$, this shows that we have rediscovered the **Euler formula**

(13)
$$e^{iy} = \cos y + i \sin y.$$

Indeed, one may use (12) for ***defining*** e^z and derive from (12) the basic properties of e^z. For instance, the differentiation formula $(e^z)' = e^z$ follows readily from (12) by termwise differentiation. ■

EXAMPLE 3 **Trigonometric and Hyperbolic Functions**

By substituting (12) into (1) of Sec. 13.6 we obtain

(14)
$$\cos z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n}}{(2n)!} = 1 - \frac{z^2}{2!} + \frac{z^4}{4!} - + \cdots$$
$$\sin z = \sum_{n=0}^{\infty} (-1)^n \frac{z^{2n+1}}{(2n+1)!} = z - \frac{z^3}{3!} + \frac{z^5}{5!} - + \cdots .$$

When $z = x$ these are the familiar Maclaurin series of the real functions $\cos x$ and $\sin x$. Similarly, by substituting (12) into (11), Sec. 13.6, we obtain

(15)
$$\cosh z = \sum_{n=0}^{\infty} \frac{z^{2n}}{(2n)!} = 1 + \frac{z^2}{2!} + \frac{z^4}{4!} + \cdots$$
$$\sinh z = \sum_{n=0}^{\infty} \frac{z^{2n+1}}{(2n+1)!} = z + \frac{z^3}{3!} + \frac{z^5}{5!} + \cdots .$$
■

EXAMPLE 4 **Logarithm**

From (1) it follows that

(16)
$$\text{Ln}\,(1 + z) = z - \frac{z^2}{2} + \frac{z^3}{3} - + \cdots \qquad (|z| < 1).$$

Replacing z by $-z$ and multiplying both sides by -1, we get

(17)
$$-\text{Ln}\,(1 - z) = \text{Ln}\,\frac{1}{1 - z} = z + \frac{z^2}{2} + \frac{z^3}{3} + \cdots \qquad (|z| < 1).$$

By adding both series we obtain

(18)
$$\text{Ln}\,\frac{1 + z}{1 - z} = 2\left(z + \frac{z^3}{3} + \frac{z^5}{5} + \cdots\right) \qquad (|z| < 1).$$
■

Practical Methods

The following examples show ways of obtaining Taylor series more quickly than by the use of the coefficient formulas. Regardless of the method used, the result will be the same. This follows from the uniqueness (see Theorem 1).

EXAMPLE 5 **Substitution**

Find the Maclaurin series of $f(z) = 1/(1 + z^2)$.

Solution. By substituting $-z^2$ for z in (11) we obtain

(19)
$$\frac{1}{1 + z^2} = \frac{1}{1 - (-z^2)} = \sum_{n=0}^{\infty} (-z^2)^n = \sum_{n=0}^{\infty} (-1)^n z^{2n} = 1 - z^2 + z^4 - z^6 + \cdots \qquad (|z| < 1).$$
■

EXAMPLE 6 **Integration**

Find the Maclaurin series of $f(z) = \arctan z$.

Solution. We have $f'(z) = 1/(1 + z^2)$. Integrating (19) term by term and using $f(0) = 0$ we get

$$\arctan z = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} z^{2n+1} = z - \frac{z^3}{3} + \frac{z^5}{5} - + \cdots \qquad (|z| < 1);$$

this series represents the principal value of $w = u + iv = \arctan z$ defined as that value for which $|u| < \pi/2$. ■

EXAMPLE 7 **Development by Using the Geometric Series**

Develop $1/(c - z)$ in powers of $z - z_0$, where $c - z_0 \neq 0$.

Solution. This was done in the proof of Theorem 1, where $c = z^*$. The beginning was simple algebra and then the use of (11) with z replaced by $(z - z_0)/(c - z_0)$:

$$\frac{1}{c-z} = \frac{1}{c - z_0 - (z - z_0)} = \frac{1}{(c - z_0)\left(1 - \dfrac{z - z_0}{c - z_0}\right)} = \frac{1}{c - z_0} \sum_{n=0}^{\infty} \left(\frac{z - z_0}{c - z_0}\right)^n$$

$$= \frac{1}{c - z_0}\left(1 + \frac{z - z_0}{c - z_0} + \left(\frac{z - z_0}{c - z_0}\right)^2 + \cdots\right).$$

This series converges for

$$\left|\frac{z - z_0}{c - z_0}\right| < 1, \qquad \text{that is,} \qquad |z - z_0| < |c - z_0|. \qquad ■$$

EXAMPLE 8 **Binomial Series, Reduction by Partial Fractions**

Find the Taylor series of the following function with center $z_0 = 1$.

$$f(z) = \frac{2z^2 + 9z + 5}{z^3 + z^2 - 8z - 12}$$

Solution. We develop $f(z)$ in partial fractions and the first fraction in a **binomial series**

$$\textbf{(20)} \qquad \frac{1}{(1+z)^m} = (1 + z)^{-m} = \sum_{n=0}^{\infty} \binom{-m}{n} z^n$$
$$= 1 - mz + \frac{m(m+1)}{2!} z^2 - \frac{m(m+1)(m+2)}{3!} z^3 + \cdots$$

with $m = 2$ and the second fraction in a geometric series, and then add the two series term by term. This gives

$$f(z) = \frac{1}{(z+2)^2} + \frac{2}{z-3} = \frac{1}{[3 + (z-1)]^2} - \frac{2}{2 - (z-1)} = \frac{1}{9}\left(\frac{1}{[1 + \frac{1}{3}(z-1)]^2}\right) - \frac{1}{1 - \frac{1}{2}(z-1)}$$

$$= \frac{1}{9}\sum_{n=0}^{\infty}\binom{-2}{n}\left(\frac{z-1}{3}\right)^n - \sum_{n=0}^{\infty}\left(\frac{z-1}{2}\right)^n = \sum_{n=0}^{\infty}\left[\frac{(-1)^n(n+1)}{3^{n+2}} - \frac{1}{2^n}\right](z-1)^n$$

$$= -\frac{8}{9} - \frac{31}{54}(z-1) - \frac{23}{108}(z-1)^2 - \frac{275}{1944}(z-1)^3 - \cdots.$$

We see that the first series converges for $|z - 1| < 3$ and the second for $|z - 1| < 2$. This had to be expected because $1/(z + 2)^2$ is singular at -2 and $2/(z - 3)$ at 3, and these points have distance 3 and 2, respectively, from the center $z_0 = 1$. Hence the whole series converges for $|z - 1| < 2$. ■

PROBLEM SET 15.4

1. **Calculus.** Which of the series in this section have you discussed in calculus? What is new?
2. **On Examples 5 and 6.** Give all the details in the derivation of the series in those examples.

3–10 MACLAURIN SERIES

Find the Maclaurin series and its radius of convergence.

3. $\sin 2z^2$
4. $\dfrac{z+2}{1-z^2}$
5. $\dfrac{1}{2+z^4}$
6. $\dfrac{1}{1+3iz}$
7. $\cos^2 \frac{1}{2}z$
8. $\sin^2 z$
9. $\displaystyle\int_0^z \exp\left(\frac{-t^2}{2}\right) dt$
10. $\displaystyle\exp(z^2)\int_0^z \exp(-t^2)\, dt$

11–14 HIGHER TRANSCENDENTAL FUNCTIONS

Find the Maclaurin series by termwise integrating the integrand. (The integrals cannot be evaluated by the usual methods of calculus. They define the **error function** erf z, **sine integral** Si(z), and **Fresnel integrals**[4] S(z) and C(z), which occur in statistics, heat conduction, optics, and other applications. These are special so-called higher transcendental functions.)

11. $\displaystyle S(z) = \int_0^z \sin t^2\, dt$
12. $\displaystyle C(z) = \int_0^z \cos t^2\, dt$
13. $\displaystyle \text{erf } z = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}\, dt$
14. $\displaystyle \text{Si}(z) = \int_0^z \frac{\sin t}{t}\, dt$

15. **CAS Project. sec, tan. (a) Euler numbers.** The Maclaurin series

$$\sec z = E_0 - \frac{E_2}{2!}z^2 + \frac{E_4}{4!}z^4 - + \cdots \tag{21}$$

defines the *Euler numbers* E_{2n}. Show that $E_0 = 1$, $E_2 = -1$, $E_4 = 5$, $E_6 = -61$. Write a program that computes the E_{2n} from the coefficient formula in (1) or extracts them as a list from the series. (For tables see Ref. [GenRef1], p. 810, listed in App. 1.)

(b) Bernoulli numbers. The Maclaurin series

$$\frac{z}{e^z - 1} = 1 + B_1 z + \frac{B_2}{2!}z^2 + \frac{B_3}{3!}z^3 + \cdots \tag{22}$$

defines the *Bernoulli numbers* B_n. Using undetermined coefficients, show that

$$\begin{aligned} &B_1 = -\tfrac{1}{2}, \quad B_2 = \tfrac{1}{6}, \quad B_3 = 0, \\ &B_4 = -\tfrac{1}{30}, \quad B_5 = 0, \quad B_6 = \tfrac{1}{42}, \cdots. \end{aligned} \tag{23}$$

Write a program for computing B_n.

(c) Tangent. Using (1), (2), Sec. 13.6, and (22), show that tan z has the following Maclaurin series and calculate from it a table of $B_0, \cdots, B_{20}$:

$$\begin{aligned} \tan z &= \frac{2i}{e^{2iz}-1} - \frac{4i}{e^{4iz}-1} - i \\ &= \sum_{n=1}^{\infty} (-1)^{n-1}\frac{2^{2n}(2^{2n}-1)}{(2n)!}B_{2n}z^{2n-1}. \end{aligned} \tag{24}$$

16. **Inverse sine.** Developing $1/\sqrt{1-z^2}$ and integrating, show that

$$\arcsin z = z + \left(\frac{1}{2}\right)\frac{z^3}{3} + \left(\frac{1\cdot 3}{2\cdot 4}\right)\frac{z^5}{5} + \left(\frac{1\cdot 3\cdot 5}{2\cdot 4\cdot 6}\right)\frac{z^7}{7} + \cdots \ (|z| < 1).$$

Show that this series represents the principal value of arcsin z (defined in Team Project 30, Sec. 13.7).

17. **TEAM PROJECT. Properties from Maclaurin Series.** Clearly, from series we can compute function values. In this project we show that properties of functions can often be discovered from their Taylor or Maclaurin series. Using suitable series, prove the following.

(a) The formulas for the derivatives of e^z, cos z, sin z, cosh z, sinh z. and Ln $(1 + z)$

(b) $\frac{1}{2}(e^{iz} + e^{-iz}) = \cos z$

(c) $\sin z \neq 0$ for all pure imaginary $z = iy \neq 0$

18–25 TAYLOR SERIES

Find the Taylor series with center z_0 and its radius of convergence.

18. $1/z, \quad z_0 = i$
19. $1/(1-z), \quad z_0 = i$
20. $\cos^2 z, \quad z_0 = \pi/2$
21. $\sin z, \quad z_0 = \pi/2$
22. $\cosh(z - \pi i), \quad z_0 = \pi i$
23. $1/(z+i)^2, \quad z_0 = i$
24. $e^{z(z-2)}, \quad z_0 = 1$
25. $\sinh(2z - i), \quad z_0 = i/2$

[4]AUGUSTIN FRESNEL (1788–1827), French physicist and engineer, known for his work in optics.

15.5 Uniform Convergence. *Optional*

We know that power series are *absolutely convergent* (Sec. 15.2, Theorem 1) and, as another basic property, we now show that they are *uniformly convergent.* Since uniform convergence is of general importance, for instance, in connection with termwise integration of series, we shall discuss it quite thoroughly.

To define uniform convergence, we consider a series whose terms are any complex functions $f_0(z), f_1(z), \cdots$

$$\sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + f_2(z) + \cdots. \tag{1}$$

(This includes power series as a special case in which $f_m(z) = a_m(z - z_0)^m$.) We assume that the series (1) converges for all z in some region G. We call its sum $s(z)$ and its nth partial sum $s_n(z)$; thus

$$s_n(z) = f_0(z) + f_1(z) + \cdots + f_n(z).$$

Convergence in G means the following. If we pick a $z = z_1$ in G, then, by the definition of convergence at z_1, for given $\epsilon > 0$ we can find an $N_1(\epsilon)$ such that

$$|s(z_1) - s_n(z_1)| < \epsilon \qquad \text{for all } n > N_1(\epsilon).$$

If we pick a z_2 in G, keeping ϵ as before, we can find an $N_2(\epsilon)$ such that

$$|s(z_2) - s_n(z_2)| < \epsilon \qquad \text{for all } n > N_2(\epsilon),$$

and so on. Hence, given an $\epsilon > 0$, to each z in G there corresponds a number $N_z(\epsilon)$. This number tells us how many terms we need (what s_n we need) at a z to make $|s(z) - s_n(z)|$ smaller than ϵ. Thus this number $N_z(\epsilon)$ measures the speed of convergence.

Small $N_z(\epsilon)$ means rapid convergence, large $N_z(\epsilon)$ means slow convergence at the point z considered. Now, if we can find an $N(\epsilon)$ larger than all these $N_z(\epsilon)$ for all z in G, we say that the convergence of the series (1) in G is *uniform.* Hence this basic concept is defined as follows.

DEFINITION

Uniform Convergence

A series (1) with sum $s(z)$ is called **uniformly convergent** in a region G if for every $\epsilon > 0$ we can find an $N = N(\epsilon)$, ***not depending on z***, such that

$$|s(z) - s_n(z)| < \epsilon \qquad \text{for all } n > N(\epsilon) \text{ \textbf{and all} } z \text{ \textbf{in} } G.$$

Uniformity of convergence is thus a property that always refers to an *infinite set* in the z-plane, that is, a set consisting of infinitely many points.

EXAMPLE 1 **Geometric Series**

Show that the geometric series $1 + z + z^2 + \cdots$ is (a) uniformly convergent in any closed disk $|z| \leqq r < 1$, (b) not uniformly convergent in its whole disk of convergence $|z| < 1$.

Solution. **(a)** For z in that closed disk we have $|1 - z| \geqq 1 - r$ (sketch it). This implies that $1/|1 - z| \leqq 1/(1 - r)$. Hence (remember (8) in Sec. 15.4 with $q = z$)

$$|s(z) - s_n(z)| = \left| \sum_{m=n+1}^{\infty} z^m \right| = \left| \frac{z^{n+1}}{1 - z} \right| \leqq \frac{r^{n+1}}{1 - r}.$$

Since $r < 1$, we can make the right side as small as we want by choosing n large enough, and since the right side does not depend on z (in the closed disk considered), this means that the convergence is uniform.

(b) For given real K (no matter how large) and n we can always find a z in the disk $|z| < 1$ such that

$$\left| \frac{z^{n+1}}{1 - z} \right| = \frac{|z|^{n+1}}{|1 - z|} > K,$$

simply by taking z close enough to 1. Hence no single $N(\epsilon)$ will suffice to make $|s(z) - s_n(z)|$ smaller than a given $\epsilon > 0$ *throughout the whole disk*. By definition, this shows that the convergence of the geometric series in $|z| < 1$ is not uniform. ■

This example suggests that *for a power series, the uniformity of convergence may at most be disturbed near the circle of convergence.* This is true:

THEOREM 1

Uniform Convergence of Power Series

A power series

$$\sum_{m=0}^{\infty} a_m (z - z_0)^m \tag{2}$$

with a nonzero radius of convergence R is uniformly convergent in every circular disk $|z - z_0| \leqq r$ of radius $r < R$.

PROOF For $|z - z_0| \leqq r$ and any positive integers n and p we have

$$|a_{n+1}(z - z_0)^{n+1} + \cdots + a_{n+p}(z - z_0)^{n+p}| \leqq |a_{n+1}|r^{n+1} + \cdots + |a_{n+p}|r^{n+p}. \tag{3}$$

Now (2) converges absolutely if $|z - z_0| = r < R$ (by Theorem 1 in Sec. 15.2). Hence it follows from the Cauchy convergence principle (Sec. 15.1) that, an $\epsilon > 0$ being given, we can find an $N(\epsilon)$ such that

$$|a_{n+1}|r^{n+1} + \cdots + |a_{n+p}|r^{n+p} < \epsilon \qquad \text{for } n > N(\epsilon) \quad \text{and} \quad p = 1, 2, \cdots.$$

From this and (3) we obtain

$$|a_{n+1}(z - z_0)^{n+1} + \cdots + a_{n+p}(z - z_0)^{n+p}| < \epsilon$$

for all z in the disk $|z - z_0| \leqq r$, every $n > N(\epsilon)$, and every $p = 1, 2, \cdots$. Since $N(\epsilon)$ is independent of z, this shows uniform convergence, and the theorem is proved. ■

Thus we have established uniform convergence of power series, the basic concern of this section. *We now shift from power series to arbitary series of variable terms* and examine uniform convergence in this more general setting. This will give a deeper understanding of uniform convergence.

Properties of Uniformly Convergent Series

Uniform convergence derives its main importance from two facts:

1. If a series of *continuous* terms is uniformly convergent, its sum is also continuous (Theorem 2, below).

2. Under the same assumptions, termwise integration is permissible (Theorem 3).

This raises two questions:

1. How can a converging series of continuous terms manage to have a discontinuous sum? (Example 2)

2. How can something go wrong in termwise integration? (Example 3)

Another natural question is:

3. What is the relation between absolute convergence and uniform convergence? The surprising answer: none. (Example 5)

These are the ideas we shall discuss.

If we add *finitely many* continuous functions, we get a continuous function as their sum. Example 2 will show that this is no longer true for an infinite series, even if it converges absolutely. However, if it converges *uniformly,* this cannot happen, as follows.

THEOREM 2

Continuity of the Sum

Let the series

$$\sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + \cdots$$

be uniformly convergent in a region G. Let $F(z)$ be its sum. Then if each term $f_m(z)$ is continuous at a point z_1 in G, the function $F(z)$ is continuous at z_1.

PROOF Let $s_n(z)$ be the nth partial sum of the series and $R_n(z)$ the corresponding remainder:

$$s_n = f_0 + f_1 + \cdots + f_n, \qquad R_n = f_{n+1} + f_{n+2} + \cdots .$$

Since the series converges uniformly, for a given $\epsilon > 0$ we can find an $N = N(\epsilon)$ such that

$$|R_N(z)| < \frac{\epsilon}{3} \qquad \text{for all } z \text{ in } G.$$

Since $s_N(z)$ is a sum of finitely many functions that are continuous at z_1, this sum is continuous at z_1. Therefore, we can find a $\delta > 0$ such that

$$|s_N(z) - s_N(z_1)| < \frac{\epsilon}{3} \quad \text{for all } z \text{ in } G \text{ for which } |z - z_1| < \delta.$$

Using $F = s_N + R_N$ and the triangle inequality (Sec. 13.2), for these z we thus obtain

$$\begin{aligned} |F(z) - F(z_1)| &= |s_N(z) + R_N(z) - [s_N(z_1) + R_N(z_1)]| \\ &\leqq |s_N(z) - s_N(z_1)| + |R_N(z)| + |R_N(z_1)| < \frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon. \end{aligned}$$

This implies that $F(z)$ is continuous at z_1, and the theorem is proved. ■

EXAMPLE 2 Series of Continuous Terms with a Discontinuous Sum

Consider the series

$$x^2 + \frac{x^2}{1+x^2} + \frac{x^2}{(1+x^2)^2} + \frac{x^2}{(1+x^2)^3} + \cdots \qquad (x \text{ real}).$$

This is a geometric series with $q = 1/(1 + x^2)$ times a factor x^2. Its nth partial sum is

$$s_n(x) = x^2 \left[1 + \frac{1}{1+x^2} + \frac{1}{(1+x^2)^2} + \cdots + \frac{1}{(1+x^2)^n}\right].$$

We now use the trick by which one finds the sum of a geometric series, namely, we multiply $s_n(x)$ by $-q = -1/(1 + x^2)$,

$$-\frac{1}{1+x^2}\, s_n(x) = -x^2 \left[\frac{1}{1+x^2} + \cdots + \frac{1}{(1+x^2)^n} + \frac{1}{(1+x^2)^{n+1}}\right].$$

Adding this to the previous formula, simplifying on the left, and canceling most terms on the right, we obtain

$$\frac{x^2}{1+x^2}\, s_n(x) = x^2 \left[1 - \frac{1}{(1+x^2)^{n+1}}\right],$$

thus

$$s_n(x) = 1 + x^2 - \frac{1}{(1+x^2)^n}.$$

The exciting Fig. 368 "explains" what is going on. We see that if $x \neq 0$, the sum is

$$s(x) = \lim_{n\to\infty} s_n(x) = 1 + x^2,$$

but for $x = 0$ we have $s_n(0) = 1 - 1 = 0$ for all n, hence $s(0) = 0$. So we have the surprising fact that the sum is discontinuous (at $x = 0$), although all the terms are continuous and the series converges even absolutely (its terms are nonnegative, thus equal to their absolute value!).

Theorem 2 now tells us that the convergence cannot be uniform in an interval containing $x = 0$. We can also verify this directly. Indeed, for $x \neq 0$ the remainder has the absolute value

$$|R_n(x)| = |s(x) - s_n(x)| = \frac{1}{(1+x^2)^n}$$

and we see that for a given ϵ (<1) we cannot find an N depending only on ϵ such that $|R_n| < \epsilon$ for all $n > N(\epsilon)$ and all x, say, in the interval $0 \leqq x \leqq 1$. ■

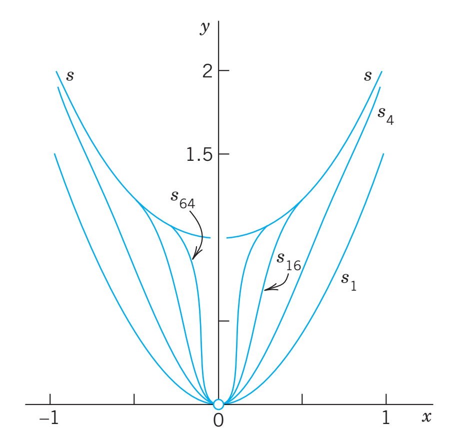


Fig. 368. Partial sums in Example 2

Termwise Integration

This is our second topic in connection with uniform convergence, and we begin with an example to become aware of the danger of just blindly integrating term-by-term.

EXAMPLE 3 **Series for Which Termwise Integration Is Not Permissible**

Let $u_m(x) = mxe^{-mx^2}$ and consider the series

$$\sum_{m=0}^{\infty} f_m(x) \qquad \text{where} \qquad f_m(x) = u_m(x) - u_{m-1}(x)$$

in the interval $0 \leqq x \leqq 1$. The nth partial sum is

$$s_n = u_1 - u_0 + u_2 - u_1 + \cdots + u_n - u_{n-1} = u_n - u_0 = u_n.$$

Hence the series has the sum $F(x) = \lim_{n\to\infty} s_n(x) = \lim_{n\to\infty} u_n(x) = 0 \quad (0 \leqq x \leqq 1)$. From this we obtain

$$\int_0^1 F(x)\, dx = 0.$$

On the other hand, by integrating term by term and using $f_1 + f_2 + \cdots + f_n = s_n$, we have

$$\sum_{m=1}^{\infty} \int_0^1 f_m(x)\, dx = \lim_{n\to\infty} \sum_{m=1}^{n} \int_0^1 f_m(x)\, dx = \lim_{n\to\infty} \int_0^1 s_n(x)\, dx.$$

Now $s_n = u_n$ and the expression on the right becomes

$$\lim_{n\to\infty} \int_0^1 u_n(x)\, dx = \lim_{n\to\infty} \int_0^1 nxe^{-nx^2}\, dx = \lim_{n\to\infty} \frac{1}{2}(1 - e^{-n}) = \frac{1}{2},$$

but not 0. This shows that the series under consideration cannot be integrated term by term from $x = 0$ to $x = 1$. ■

The series in Example 3 is not uniformly convergent in the interval of integration, and we shall now prove that in the case of a uniformly convergent series of continuous functions we may integrate term by term.

THEOREM 3 **Termwise Integration**

Let

$$F(z) = \sum_{m=0}^{\infty} f_m(z) = f_0(z) + f_1(z) + \cdots$$

be a uniformly convergent series of continuous functions in a region G. Let C be any path in G. Then the series

$$\sum_{m=0}^{\infty} \int_C f_m(z)\, dz = \int_C f_0(z)\, dz + \int_C f_1(z)\, dz + \cdots \tag{4}$$

is convergent and has the sum $\int_C F(z)\, dz$.

PROOF From Theorem 2 it follows that $F(z)$ is continuous. Let $s_n(z)$ be the nth partial sum of the given series and $R_n(z)$ the corresponding remainder. Then $F = s_n + R_n$ and by integration,

$$\int_C F(z)\, dz = \int_C s_n(z)\, dz + \int_C R_n(z)\, dz.$$

Let L be the length of C. Since the given series converges uniformly, for every given $\epsilon > 0$ we can find a number N such that $|R_n(z)| < \epsilon/L$ for all $n > N$ and all z in G. By applying the ML-inequality (Sec. 14.1) we thus obtain

$$\left| \int_C R_n(z)\, dz \right| < \frac{\epsilon}{L} L = \epsilon \qquad \text{for all } n > N.$$

Since $R_n = F - s_n$, this means that

$$\left| \int_C F(z)\, dz - \int_C s_n(z)\, dz \right| < \epsilon \qquad \text{for all } n > N.$$

Hence, the series (4) converges and has the sum indicated in the theorem. ■

Theorems 2 and 3 characterize the two most important properties of uniformly convergent series. Also, since differentiation and integration are inverse processes, Theorem 3 implies

THEOREM 4

Termwise Differentiation

Let the series $f_0(z) + f_1(z) + f_2(z) + \cdots$ be convergent in a region G and let $F(z)$ be its sum. Suppose that the series $f_0'(z) + f_1'(z) + f_2'(z) + \cdots$ converges uniformly in G and its terms are continuous in G. Then

$$F'(z) = f_0'(z) + f_1'(z) + f_2'(z) + \cdots \qquad \textit{for all } z \textit{ in } G.$$

Test for Uniform Convergence

Uniform convergence is usually proved by the following comparison test.

THEOREM 5

Weierstrass[5] *M*-Test for Uniform Convergence

Consider a series of the form (1) *in a region G of the z-plane. Suppose that one can find a convergent series of constant terms,*

$$M_0 + M_1 + M_2 + \cdots, \tag{5}$$

such that $|f_m(z)| \leqq M_m$ for all z in G and every $m = 0, 1, \cdots$. Then (1) *is uniformly convergent in G.*

The simple proof is left to the student (Team Project 18).

[5]KARL WEIERSTRASS (1815–1897), great German mathematician, who developed complex analysis based on the concept of power series and residue integration. (See footnote in Section 13.4.) He put analysis on a sound theoretical footing. His mathematical rigor is so legendary that one speaks *Weierstrassian rigor*. (See paper by Birkhoff and Kreyszig, 1984 in footnote in Sec. 5.5; Kreyszig, E., On the Calculus, of Variations and Its Major Influences on the Mathematics of the First Half of Our Century. Part II, *American Mathematical Monthly* (1994), 101, No. 9, pp. 902–908). Weierstrass also made contributions to the calculus of variations, approximation theory, and differential geometry. He obtained the concept of uniform convergence in 1841 (published 1894, *sic!*); the first publication on the concept was by G. G. STOKES (see Sec 10.9) in 1847.

EXAMPLE 4 Weierstrass *M*-Test

Does the following series converge uniformly in the disk $|z| \leqq 1$?

$$\sum_{m=1}^{\infty} \frac{z^m + 1}{m^2 + \cosh m|z|}.$$

Solution. Uniform convergence follows by the Weierstrass *M*-test and the convergence of $\Sigma 1/m^2$ (see Sec. 15.1, in the proof of Theorem 8) because

$$\left| \frac{z^m + 1}{m^2 + \cosh m|z|} \right| \leqq \frac{|z|^m + 1}{m^2} \leqq \frac{2}{m^2}.$$

■

No Relation Between Absolute and Uniform Convergence

We finally show the surprising fact that there are series that converge absolutely but not uniformly, and others that converge uniformly but not absolutely, so that there is no relation between the two concepts.

EXAMPLE 5 No Relation Between Absolute and Uniform Convergence

The series in Example 2 converges absolutely but not uniformly, as we have shown. On the other hand, the series

$$\sum_{m=1}^{\infty} \frac{(-1)^{m-1}}{x^2 + m} = \frac{1}{x^2 + 1} - \frac{1}{x^2 + 2} + \frac{1}{x^2 + 3} - + \cdots \qquad (x \text{ real})$$

converges uniformly on the whole real line but not absolutely.

Proof. By the familiar Leibniz test of calculus (see App. A3.3) the remainder R_n does not exceed its first term in absolute value, since we have a series of alternating terms whose absolute values form a monotone decreasing sequence with limit zero. Hence given $\epsilon > 0$, for all x we have

$$|R_n(x)| \leqq \frac{1}{x^2 + n + 1} < \frac{1}{n} < \epsilon \qquad \text{if } n > N(\epsilon) \geqq \frac{1}{\epsilon}.$$

This proves uniform convergence, since $N(\epsilon)$ does not depend on x.

The convergence is not absolute because for any *fixed* x we have

$$\left| \frac{(-1)^{m-1}}{x^2 + m} \right| = \frac{1}{x^2 + m} > \frac{k}{m}$$

where k is a suitable constant, and $k\Sigma 1/m$ diverges. ■

PROBLEM SET 15.5

1. **CAS EXPERIMENT. Graphs of Partial Sums. (a) Fig. 368.** Produce this exciting figure using your CAS. Add further curves, say, those of s_{256}, s_{1024}, etc. on the same screen.

(b) Power series. Study the nonuniformity of convergence experimentally by graphing partial sums near the endpoints of the convergence interval for real $z = x$.

2–9 POWER SERIES

Where does the power series converge uniformly? Give reason.

2. $\sum_{n=0}^{\infty} \left(\frac{n+2}{7n-3}\right)^n z^n$

3. $\sum_{n=0}^{\infty} \frac{1}{3^n}(z+i)^{2n}$

4. $\sum_{n=0}^{\infty} \frac{3^n(1-i)^n}{n!}(z-i)^n$

5. $\sum_{n=2}^{\infty} \binom{n}{2}(4z+2i)^n$

6. $\sum_{n=0}^{\infty} 2^n(\tanh n^2)z^{2n}$

7. $\sum_{n=1}^{\infty} \frac{n!}{n^2}\left(z+\frac{1}{2}i\right)$

8. $\sum_{n=1}^{\infty} \frac{3^n}{n(n+1)}(z-1)^{2n}$

9. $\sum_{n=1}^{\infty} \frac{(-1)^n}{2^n n^2}(z-2i)^n$

10–17 UNIFORM CONVERGENCE

Prove that the series converges uniformly in the indicated region.

10. $\sum_{n=0}^{\infty} \frac{z^{2n}}{2n!}, \quad |z| \leqq 10^{20}$

11. $\sum_{n=1}^{\infty} \frac{z^n}{n^2}, \quad |z| \leqq 1$

12. $\sum_{n=1}^{\infty} \frac{z^n}{n^3 \cosh n|z|}, \quad |z| \leqq 1$

13. $\sum_{n=1}^{\infty} \frac{\sin^n |z|}{n^2}, \quad$ all z

14. $\sum_{n=0}^{\infty} \frac{z^n}{|z|^{2n}+1}, \quad 2 \leqq |z| \leqq 10$

15. $\sum_{n=0}^{\infty} \frac{(n!)^2}{(2n!)}z^n, \quad |z| \leqq 3$

16. $\sum_{n=1}^{\infty} \frac{\tanh^n |z|}{n(n+1)}, \quad$ all z

17. $\sum_{n=1}^{\infty} \frac{\pi^n}{n^4}z^{2n}, \quad |z| \leqq 0.56$

18. **TEAM PROJECT. Uniform Convergence.**

(a) Weierstrass *M*-test. Give a proof.

(b) Termwise differentiation. Derive Theorem 4 from Theorem 3.

(c) Subregions. Prove that uniform convergence of a series in a region G implies uniform convergence in any portion of G. Is the converse true?

(d) Example 2. Find the precise region of convergence of the series in Example 2 with x replaced by a complex variable z.

(e) Figure 369. Show that $x^2 \sum_{m=1}^{\infty} (1+x^2)^{-m} = 1$ if $x \neq 0$ and 0 if $x = 0$. Verify by computation that the partial sums s_1, s_2, s_3 look as shown in Fig. 369.

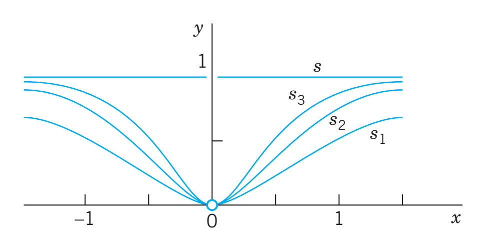


Fig. 369. Sum s and partial sums in Team Project 18(e)

19–20 HEAT EQUATION

Show that (9) in Sec. 12.6 with coefficients (10) is a solution of the heat equation for $t > 0$, assuming that $f(x)$ is continuous on the interval $0 \leqq x \leqq L$ and has one-sided derivatives at all interior points of that interval. Proceed as follows.

19. Show that $|B_n|$ is bounded, say $|B_n| < K$ for all n. Conclude that

$$|u_n| < Ke^{-\lambda_n^2 t_0} \quad \text{if} \quad t \geqq t_0 > 0$$

and, by the Weierstrass test, the series (9) converges uniformly with respect to x and t for $t \geqq t_0$, $0 \leqq x \leqq L$. Using Theorem 2, show that $u(x, t)$ is continuous for $t \geqq t_0$ and thus satisfies the boundary conditions (2) for $t \geqq t_0$.

20. Show that $|\partial u_n/\partial t| < \lambda_n^2 K e^{-\lambda_n^2 t_0}$ if $t \geqq t_0$ and the series of the expressions on the right converges, by the ratio test. Conclude from this, the Weierstrass test, and Theorem 4 that the series (9) can be differentiated term by term with respect to t and the resulting series has the sum $\partial u/\partial t$. Show that (9) can be differentiated twice with respect to x and the resulting series has the sum $\partial^2 u/\partial x^2$. Conclude from this and the result to Prob. 19 that (9) is a solution of the heat equation for all $t \geqq t_0$. (The proof that (9) satisfies the given initial condition can be found in Ref. [C10] listed in App. 1.)

CHAPTER 15 REVIEW QUESTIONS AND PROBLEMS

1. What is convergence test for series? State two tests from memory. Give examples.

2. What is a power series? Why are these series very important in complex analysis?

3. What is absolute convergence? Conditional convergence? Uniform convergence?

4. What do you know about convergence of power series?

5. What is a Taylor series? Give some basic examples.

6. What do you know about adding and multiplying power series?

7. Does every function have a Taylor series development? Explain.

8. Can properties of functions be discovered from Maclaurin series? Give examples.

9. What do you know about termwise integration of series?

10. How did we obtain Taylor's formula from Cauchy's formula?

11–15 RADIUS OF CONVERGENCE

Find the radius of convergence.

11. $\sum_{n=2}^{\infty} \frac{n+1}{n^2+1}(z+1)^n$

12. $\sum_{n=2}^{\infty} \frac{4^n}{n-1}(z-\pi i)^n$

13. $\sum_{n=2}^{\infty} \frac{n(n-1)}{3^n}(z-i)^n$

14. $\sum_{n=1}^{\infty} \frac{n^5}{n!}(z-3i)^{2n}$

15. $\sum_{n=1}^{\infty} \frac{(-2)^{n+1}}{2n} z^n$

16–20 RADIUS OF CONVERGENCE

Find the radius of convergence. Try to identify the sum of the series as a familiar function.

16. $\sum_{n=1}^{\infty} \frac{z^n}{n}$

17. $\sum_{n=0}^{\infty} \frac{z^n}{n!} z^n$

18. $\sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!}(\pi z)^{2n+1}$

19. $\sum_{n=0}^{\infty} \frac{z^n}{(2n)!}$

20. $\sum_{n=0}^{\infty} \frac{z^n}{(3+4i)^n}$

21–25 MACLAURIN SERIES

Find the Maclaurin series and its radius of convergence. Show details.

21. $(\sinh z^2)/z^2$

22. $1/(1-z)^3$

23. $\cos^2 z$

24. $1/(\pi z+1)$

25. $-(\exp/(-z^2)-1)/z^2$

26–30 TAYLOR SERIES

Find the Taylor series with the given point as center and its radius of convergence.

26. z^4, $\quad i$

27. $\cos z$, $\quad \frac{1}{2}\pi$

28. $1/z$, $\quad 2i$

29. $\operatorname{Ln} z$, $\quad 3$

30. e^z, $\quad \pi i$

SUMMARY OF CHAPTER 15
Power Series, Taylor Series

Sequences, series, and convergence tests are discussed in Sec. 15.1. A **power series** is of the form (Sec. 15.2)

$$\sum_{n=0}^{\infty} a_n(z-z_0)^n = a_0 + a_1(z-z_0) + a_2(z-z_0)^2 + \cdots; \tag{1}$$

z_0 is its *center*. The series (1) converges for $|z-z_0| < R$ and diverges for $|z-z_0| > R$, where R is the **radius of convergence**. Some power series converge

for all z (then we write $R = \infty$). In exceptional cases a power series may converge only at the center; such a series is practically useless. Also, $R = \lim |a_n/a_{n+1}|$ if this limit exists. The series (1) converges absolutely (Sec. 15.2) and ***uniformly*** (Sec. 15.5) in every closed disk $|z - z_0| \leqq r < R$ $(R > 0)$. It represents an analytic function $f(z)$ for $|z - z_0| < R$. The derivatives $f'(z), f''(z), \cdots$ are obtained by termwise differentiation of (1), and these series have the same radius of convergence R as (1). See Sec. 15.3.

Conversely, *every* analytic function $f(z)$ can be represented by power series. These **Taylor series** of $f(z)$ are of the form (Sec. 15.4)

$$f(z) = \sum_{n=0}^{\infty} \frac{1}{n!} f^{(n)}(z_0)(z - z_0)^n \qquad (|z - z_0| < R), \tag{2}$$

as in calculus. They converge for all z in the open disk with center z_0 and radius generally equal to the distance from z_0 to the nearest **singularity** of $f(z)$ (point at which $f(z)$ ceases to be analytic as defined in Sec. 15.4). If $f(z)$ is **entire** (analytic for all z; see Sec. 13.5), then (2) converges for all z. The functions e^z, $\cos z$, $\sin z$, etc. have Maclaurin series, that is, Taylor series with center 0, similar to those in calculus (Sec. 15.4).

CHAPTER 16

Laurent Series. Residue Integration

The main purpose of this chapter is to learn about another powerful method for evaluating complex integrals and certain real integrals. It is called *residue integration*. Recall that the first method of evaluating complex integrals consisted of directly applying Cauchy's integral formula of Sec. 14.3. Then we learned about Taylor series (Chap. 15) and will now generalize Taylor series. The beauty of residue integration, the second method of integration, is that it brings together a lot of the previous material.

Laurent series generalize Taylor series. Indeed, whereas a Taylor series has positive integer powers (and a constant term) and converges in a disk, a ***Laurent series*** (Sec. 16.1) is a series of positive ***and negative*** integer powers of $z - z_0$ and converges in an annulus (a circular ring) with center z_0. Hence, by a Laurent series, we can represent a given function $f(z)$ that is analytic in an annulus and may have singularities outside the ring as well as in the "hole" of the annulus.

We know that for a given function the Taylor series with a given center z_0 is unique. We shall see that, in contrast, a function $f(z)$ can have several Laurent series with the same center z_0 and valid in several concentric annuli. The most important of these series is the one that converges for $0 < |z - z_0| < R$, that is, everywhere near the center z_0 except at z_0 itself, where z_0 is a singular point of $f(z)$. The series (or finite sum) of the negative powers of *this* Laurent series is called the **principal part** of the singularity of $f(z)$ at z_0, and is used to classify this singularity (Sec. 16.2). The coefficient of the power $1/(z - z_0)$ of *this* series is called the **residue** of $f(z)$ at z_0. Residues are used in an elegant and powerful integration method, called ***residue integration***, for complex contour integrals (Sec. 16.3) as well as for certain complicated real integrals (Sec. 16.4).

Prerequisite: Chaps. 13, 14, Sec. 15.2.
Sections that may be omitted in a shorter course: 16.2, 16.4.
References and Answers to Problems: App. 1 Part D, App. 2.

16.1 Laurent Series

Laurent series generalize Taylor series. If, in an application, we want to develop a function $f(z)$ in powers of $z - z_0$ when $f(z)$ is singular at z_0 (as defined in Sec. 15.4), we cannot use a Taylor series. Instead we can use a new kind of series, called **Laurent series**,[1]

[1] PIERRE ALPHONSE LAURENT (1813–1854), French military engineer and mathematician, published the theorem in 1843.

consisting of positive integer powers of $z - z_0$ (and a constant) as well as ***negative integer powers*** of $z - z_0$; this is the new feature.

Laurent series are also used for classifying singularities (Sec. 16.2) and in a powerful integration method ("residue integration," Sec. 16.3).

A Laurent series of $f(z)$ converges in an annulus (in the "hole" of which $f(z)$ may have singularities), as follows.

THEOREM 1

Laurent's Theorem

Let $f(z)$ be analytic in a domain containing two concentric circles C_1 and C_2 with center z_0 and the annulus between them (blue in Fig. 370). *Then $f(z)$ can be represented by the Laurent series*

$$
\begin{aligned}
f(z) &= \sum_{n=0}^{\infty} a_n (z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \\
&= a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots \\
&\qquad \cdots + \frac{b_1}{z - z_0} + \frac{b_2}{(z - z_0)^2} + \cdots
\end{aligned} \tag{1}
$$

consisting of nonnegative and negative powers. The coefficients of this Laurent series are given by the integrals

$$
a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^*, \qquad b_n = \frac{1}{2\pi i} \oint_C (z^* - z_0)^{n-1} f(z^*)\, dz^*, \tag{2}
$$

taken counterclockwise around any simple closed path C that lies in the annulus and encircles the inner circle, as in Fig. 370. [The variable of integration is denoted by z^* since z is used in (1).]

This series converges and represents $f(z)$ in the enlarged open annulus obtained from the given annulus by continuously increasing the outer circle C_1 and decreasing C_2 until each of the two circles reaches a point where $f(z)$ is singular.

In the important special case that z_0 is the only singular point of $f(z)$ inside C_2, this circle can be shrunk to the point z_0, giving convergence in a disk except at the center. In this case the series (or finite sum) of the negative powers of (1) *is called the* **principal part** *of $f(z)$ at z_0* [*or of that Laurent series* (1)].

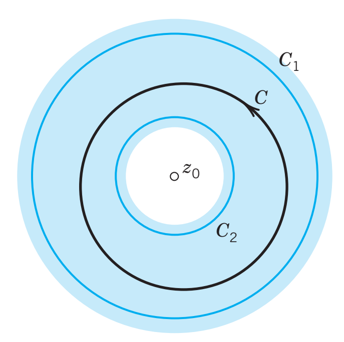


Fig. 370. Laurent's theorem

COMMENT. Obviously, instead of (1), (2) we may write (denoting b_n by a_{-n})

$$f(z) = \sum_{n=-\infty}^{\infty} a_n(z - z_0)^n \tag{1'}$$

where all the coefficients are now given by a single integral formula, namely,

$$a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^* \qquad (n = 0, \pm 1, \pm 2, \cdots). \tag{2'}$$

Let us now prove Laurent's theorem.

PROOF **(a)** ***The nonnegative powers*** are those of a Taylor series.

To see this, we use Cauchy's integral formula (3) in Sec. 14.3 with z^* (instead of z) as the variable of integration and z instead of z_0. Let $g(z)$ and $h(z)$ denote the functions represented by the two terms in (3), Sec. 14.3. Then

$$f(z) = g(z) + h(z) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{z^* - z}\, dz^* - \frac{1}{2\pi i} \oint_{C_2} \frac{f(z^*)}{z^* - z}\, dz^*. \tag{3}$$

Here z is any point in the given annulus and we integrate counterclockwise over both C_1 and C_2, so that the minus sign appears since in (3) of Sec. 14.3 the integration over C_2 is taken clockwise. We transform each of these two integrals as in Sec. 15.4. The first integral is precisely as in Sec. 15.4. Hence we get exactly the same result, namely, the Taylor series of $g(z)$,

$$g(z) = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{z^* - z}\, dz^* = \sum_{n=0}^{\infty} a_n(z - z_0)^n \tag{4}$$

with coefficients [see (2), Sec. 15.4, counterclockwise integration]

$$a_n = \frac{1}{2\pi i} \oint_{C_1} \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^*. \tag{5}$$

Here we can replace C_1 by C (see Fig. 370), by the principle of deformation of path, since z_0, the point where the integrand in (5) is not analytic, is not a point of the annulus. This proves the formula for the a_n in (2).

(b) ***The negative powers*** in (1) and the formula for b_n in (2) are obtained if we consider $h(z)$. It consists of the second integral times $-1/(2\pi i)$ in (3). Since z lies in the annulus, it lies in the exterior of the path C_2. Hence the situation differs from that for the first integral. The essential point is that instead of [see (7*) in Sec. 15.4]

$$\text{(6)} \qquad \text{(a)} \quad \left|\frac{z - z_0}{z^* - z_0}\right| < 1 \qquad \text{we now have} \qquad \text{(b)} \quad \left|\frac{z^* - z_0}{z - z_0}\right| < 1.$$

Consequently, we must develop the expression $1/(z^* - z)$ in the integrand of the second integral in (3) in powers of $(z^* - z_0)/(z - z_0)$ (instead of the reciprocal of this) to get a *convergent* series. We find

$$\frac{1}{z^* - z} = \frac{1}{z^* - z_0 - (z - z_0)} = \frac{-1}{(z - z_0)\left(1 - \dfrac{z^* - z_0}{z - z_0}\right)}.$$

Compare this for a moment with (7) in Sec. 15.4, to really understand the difference. Then go on and apply formula (8), Sec. 15.4, for a finite geometric sum, obtaining

$$\begin{aligned}\frac{1}{z^* - z} = -\frac{1}{z - z_0}&\left\{1 + \frac{z^* - z_0}{z - z_0} + \left(\frac{z^* - z_0}{z - z_0}\right)^2 + \cdots + \left(\frac{z^* - z_0}{z - z_0}\right)^n\right\}\\ &- \frac{1}{z - z^*}\left(\frac{z^* - z_0}{z - z_0}\right)^{n+1}.\end{aligned}$$

Multiplication by $-f(z^*)/2\pi i$ and integration over C_2 on both sides now yield

$$\begin{aligned}h(z) &= -\frac{1}{2\pi i}\oint_{C_2}\frac{f(z^*)}{z^* - z}\,dz^*\\ &= \frac{1}{2\pi i}\left\{\frac{1}{z - z_0}\oint_{C_2} f(z^*)\,dz^* + \frac{1}{(z - z_0)^2}\oint_{C_2}(z^* - z_0)\,f(z^*)\,dz^* + \cdots\right.\\ &\qquad + \frac{1}{(z - z_0)^n}\oint_{C_2}(z^* - z_0)^{n-1} f(z^*)\,dz^*\\ &\qquad\qquad \left. + \frac{1}{(z - z_0)^{n+1}}\oint_{C_2}(z^* - z_0)^n f(z^*)\,dz^*\right\} + R_n^*(z)\end{aligned}$$

with the last term on the right given by

$$\text{(7)} \qquad R_n^*(z) = \frac{1}{2\pi i(z - z_0)^{n+1}}\oint_{C_2}\frac{(z^* - z_0)^{n+1}}{z - z^*}f(z^*)\,dz^*.$$

As before, we can integrate over C instead of C_2 in the integrals on the right. We see that on the right, the power $1/(z - z_0)^n$ is multiplied by b_n as given in (2). This establishes Laurent's theorem, provided

$$\text{(8)} \qquad \lim_{n\to\infty} R_n^*(z) = 0.$$

(c) ***Convergence proof of* (8).** Very often (1) will have only finitely many negative powers. Then there is nothing to be proved. Otherwise, we begin by noting that $f(z^*)/(z - z^*)$ in (7) is bounded in absolute value, say,

$$\left|\frac{f(z^*)}{z - z^*}\right| < \widetilde{M} \qquad \text{for all } z^* \text{ on } C_2$$

because $f(z^*)$ is analytic in the annulus and on C_2, and z^* lies on C_2 and z outside, so that $z - z^* \neq 0$. From this and the ML-inequality (Sec. 14.1) applied to (7) we get the inequality ($L = 2\pi r_2$ = length of C_2, $r_2 = |z^* - z_0|$ = radius of C_2 = const)

$$|R_n^*(z)| \leqq \frac{1}{2\pi|z - z_0|^{n+1}} r_2^{n+1} \widetilde{M}L = \frac{\widetilde{M}L}{2\pi}\left(\frac{r_2}{|z - z_0|}\right)^{n+1}.$$

From (6b) we see that the expression on the right approaches zero as n approaches infinity. This proves (8). The representation (1) with coefficients (2) is now established in the given annulus.

(d) ***Convergence of*** **(1)** ***in the enlarged annulus.*** The first series in (1) is a Taylor series [representing $g(z)$]; hence it converges in the disk D with center z_0 whose radius equals the distance of the singularity (or singularities) closest to z_0. Also, $g(z)$ must be singular at all points outside C_1 where $f(z)$ is singular.

The second series in (1), representing $h(z)$, is a power series in $Z = 1/(z - z_0)$. Let the given annulus be $r_2 < |z - z_0| < r_1$, where r_1 and r_2 are the radii of C_1 and C_2, respectively (Fig. 370). This corresponds to $1/r_2 > |Z| > 1/r_1$. Hence this power series in Z must converge at least in the disk $|Z| < 1/r_2$. This corresponds to the exterior $|z - z_0| > r_2$ of C_2, so that $h(z)$ is analytic for all z outside C_2. Also, $h(z)$ must be singular inside C_2 where $f(z)$ is singular, and the series of the negative powers of (1) converges for all z in the exterior E of the circle with center z_0 and radius equal to the maximum distance from z_0 to the singularities of $f(z)$ inside C_2. The domain common to D and E is the enlarged open annulus characterized near the end of Laurent's theorem, whose proof is now complete. ■

Uniqueness. *The Laurent series of a given analytic function $f(z)$ in its annulus of convergence is unique* (see Team Project 18). *However, $f(z)$ may have different Laurent series in two annuli with the same center*; see the examples below. The uniqueness is essential. As for a Taylor series, to obtain the coefficients of Laurent series, we do not generally use the integral formulas (2); instead, we use various other methods, some of which we shall illustrate in our examples. If a Laurent series has been found by any such process, the uniqueness guarantees that it must be ***the*** Laurent series of the given function in the given annulus.

EXAMPLE 1 **Use of Maclaurin Series**

Find the Laurent series of $z^{-5} \sin z$ with center 0.

Solution. By (14), Sec. 15.4, we obtain

$$z^{-5} \sin z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n-4} = \frac{1}{z^4} - \frac{1}{6z^2} + \frac{1}{120} - \frac{1}{5040} z^2 + - \cdots \qquad (|z| > 0).$$

Here the "annulus" of convergence is the whole complex plane without the origin and the principal part of the series at 0 is $z^{-4} - \frac{1}{6} z^{-2}$. ■

EXAMPLE 2 **Substitution**

Find the Laurent series of $z^2 e^{1/z}$ with center 0.

Solution. From (12) in Sec. 15.4 with z replaced by $1/z$ we obtain a Laurent series whose principal part is an infinite series,

$$z^2 e^{1/z} = z^2\left(1 + \frac{1}{1!z} + \frac{1}{2!z^2} + \cdots\right) = z^2 + z + \frac{1}{2} + \frac{1}{3!z} + \frac{1}{4!z^2} + \cdots \qquad (|z| > 0).$$ ■

EXAMPLE 3 **Development of $1/(1-z)$**

Develop $1/(1-z)$ **(a)** in nonnegative powers of z, **(b)** in negative powers of z.

Solution.

(a) $$\frac{1}{1-z} = \sum_{n=0}^{\infty} z^n \qquad \text{(valid if } |z| < 1\text{).}$$

(b) $$\frac{1}{1-z} = \frac{-1}{z(1-z^{-1})} = -\sum_{n=0}^{\infty} \frac{1}{z^{n+1}} = -\frac{1}{z} - \frac{1}{z^2} - \cdots \qquad \text{(valid if } |z| > 1\text{).}$$ ■

EXAMPLE 4 **Laurent Expansions in Different Concentric Annuli**

Find all Laurent series of $1/(z^3 - z^4)$ with center 0.

Solution. Multiplying by $1/z^3$, we get from Example 3

(I) $$\frac{1}{z^3 - z^4} = \sum_{n=0}^{\infty} z^{n-3} = \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + \cdots \qquad (0 < |z| < 1),$$

(II) $$\frac{1}{z^3 - z^4} = -\sum_{n=0}^{\infty} \frac{1}{z^{n+4}} = -\frac{1}{z^4} - \frac{1}{z^5} - \cdots \qquad (|z| > 1).$$ ■

EXAMPLE 5 **Use of Partial Fractions**

Find all Taylor and Laurent series of $f(z) = \dfrac{-2z + 3}{z^2 - 3z + 2}$ with center 0.

Solution. In terms of partial fractions,

$$f(z) = -\frac{1}{z-1} - \frac{1}{z-2}.$$

(a) and (b) in Example 3 take care of the first fraction. For the second fraction,

(c) $$-\frac{1}{z-2} = \frac{1}{2\left(1 - \frac{1}{2}z\right)} = \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} z^n \qquad (|z| < 2),$$

(d) $$-\frac{1}{z-2} = -\frac{1}{z\left(1 - \frac{2}{z}\right)} = -\sum_{n=0}^{\infty} \frac{2^n}{z^{n+1}} \qquad (|z| > 2).$$

(I) From (a) and (c), valid for $|z| < 1$ (see Fig. 371),

$$f(z) = \sum_{n=0}^{\infty} \left(1 + \frac{1}{2^{n+1}}\right) z^n = \frac{3}{2} + \frac{5}{4}z + \frac{9}{8}z^2 + \cdots.$$

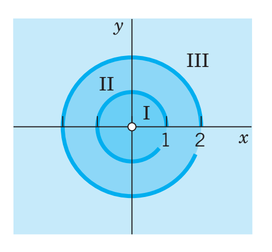


Fig. 371. Regions of convergence in Example 5

(II) From (c) and (b), valid for $1 < |z| < 2$,

$$f(z) = \sum_{n=0}^{\infty} \frac{1}{2^{n+1}} z^n - \sum_{n=0}^{\infty} \frac{1}{z^{n+1}} = \frac{1}{2} + \frac{1}{4}z + \frac{1}{8}z^2 + \cdots - \frac{1}{z} - \frac{1}{z^2} - \cdots.$$

(III) From (d) and (b), valid for $|z| > 2$,

$$f(z) = -\sum_{n=0}^{\infty} (2^n + 1) \frac{1}{z^{n+1}} = -\frac{2}{z} - \frac{3}{z^2} - \frac{5}{z^3} - \frac{9}{z^4} - \cdots.$$ ■

If $f(z)$ in Laurent's theorem is analytic inside C_2, the coefficients b_n in (2) are zero by Cauchy's integral theorem, so that the Laurent series reduces to a Taylor series. Examples 3(a) and 5(I) illustrate this.

PROBLEM SET 16.1

1–8 LAURENT SERIES NEAR A SINGULARITY AT 0

Expand the function in a Laurent series that converges for $0 < |z| < R$ and determine the precise region of convergence. Show the details of your work.

1. $\dfrac{\cos z}{z^4}$

2. $\dfrac{\exp(-1/z^2)}{z^2}$

3. $\dfrac{\exp z^2}{z^3}$

4. $\dfrac{\sin \pi z}{z^2}$

5. $\dfrac{1}{z^2 - z^3}$

6. $\dfrac{\sinh 2z}{z^2}$

7. $z^3 \cosh \dfrac{1}{z}$

8. $\dfrac{e^z}{z^2 - z^3}$

9–16 LAURENT SERIES NEAR A SINGULARITY AT z_0

Find the Laurent series that converges for $0 < |z - z_0| < R$ and determine the precise region of convergence. Show details.

9. $\dfrac{e^z}{(z-1)^2}, \quad z_0 = 1$

10. $\dfrac{z^2 - 3i}{(z-3)^2}, \quad z_0 = 3$

11. $\dfrac{z^2}{(z - \pi i)^4}, \quad z_0 = \pi i$

12. $\dfrac{1}{z^2(z-i)}, \quad z_0 = i$

13. $\dfrac{1}{z^3(z-i)^2}, \quad z_0 = i$

14. $\dfrac{e^{az}}{z-b}, \quad z_0 = b$

15. $\dfrac{\cos z}{(z-\pi)^2}, \quad z_0 = \pi$

16. $\dfrac{\sin z}{(z - \frac{1}{4}\pi)^3}, \quad z_0 = \frac{1}{4}\pi$

17. CAS PROJECT. Partial Fractions. Write a program for obtaining Laurent series by the use of partial fractions. Using the program, verify the calculations in Example 5 of the text. Apply the program to two other functions of your choice.

18. TEAM PROJECT. Laurent Series. (a) Uniqueness. Prove that the Laurent expansion of a given analytic function in a given annulus is unique.

(b) Accumulation of singularities. Does $\tan(1/z)$ have a Laurent series that converges in a region $0 < |z| < R$? (Give a reason.)

(c) Integrals. Expand the following functions in a Laurent series that converges for $|z| > 0$:

$$\frac{1}{z^2}\int_0^z \frac{e^t - 1}{t}\,dt, \qquad \frac{1}{z^3}\int_0^z \frac{\sin t}{t}\,dt.$$

19–25 TAYLOR AND LAURENT SERIES

Find all Taylor and Laurent series with center z_0. Determine the precise regions of convergence. Show details.

19. $\dfrac{1}{1-z^2}, \quad z_0 = 0$

20. $\dfrac{1}{z}, \quad z_0 = 1$

21. $\dfrac{\sin z}{z + \frac{1}{2}\pi}, \quad z_0 = -\frac{1}{2}\pi$

22. $\dfrac{1}{z^2}, \quad z_0 = i$

23. $\dfrac{z^8}{1-z^4}, \quad z_0 = 0$

24. $\dfrac{\sinh z}{(z-1)^4}, \quad z_0 = 1$

25. $\dfrac{z^3 - 2iz^2}{(z-i)^2}, \quad z_0 = i$

16.2 Singularities and Zeros. Infinity

Roughly, a *singular point* of an analytic function $f(z)$ is a z_0 at which $f(z)$ ceases to be analytic, and a *zero* is a z at which $f(z) = 0$. Precise definitions follow below. In this section we show that Laurent series can be used for classifying singularities and Taylor series for discussing zeros.

Singularities were defined in Sec. 15.4, as we shall now recall and extend. We also remember that, by definition, a function is a ***single-valued*** relation, as was emphasized in Sec. 13.3.

We say that a function $f(z)$ **is singular** or **has a singularity** at a point $z = z_0$ if $f(z)$ is not analytic (perhaps not even defined) at $z = z_0$, but every neighborhood of $z = z_0$ contains points at which $f(z)$ is analytic. We also say that $z = z_0$ is a **singular point** of $f(z)$.

We call $z = z_0$ an **isolated singularity** of $f(z)$ if $z = z_0$ has a neighborhood without further singularities of $f(z)$. *Example:* $\tan z$ has isolated singularities at $\pm\pi/2, \pm 3\pi/2$, etc.; $\tan(1/z)$ has a nonisolated singularity at 0. (Explain!)

Isolated singularities of $f(z)$ at $z = z_0$ can be classified by the Laurent series

$$f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \qquad \text{(Sec. 16.1)} \tag{1}$$

valid ***in the immediate neighborhood*** of the singular point $z = z_0$, except at z_0 itself, that is, in a region of the form

$$0 < |z - z_0| < R.$$

The sum of the first series is analytic at $z = z_0$, as we know from the last section. The second series, containing the negative powers, is called the **principal part** of (1), as we remember from the last section. If it has only finitely many terms, it is of the form

$$\frac{b_1}{z - z_0} + \cdots + \frac{b_m}{(z - z_0)^m} \qquad (b_m \neq 0). \tag{2}$$

Then the singularity of $f(z)$ at $z = z_0$ is called a **pole**, and m is called its **order**. Poles of the first order are also known as **simple poles**.

If the principal part of (1) has infinitely many terms, we say that $f(z)$ has at $z = z_0$ an **isolated essential singularity**.

We leave aside nonisolated singularities.

EXAMPLE 1 Poles. Essential Singularities

The function

$$f(z) = \frac{1}{z(z - 2)^5} + \frac{3}{(z - 2)^2}$$

has a simple pole at $z = 0$ and a pole of fifth order at $z = 2$. Examples of functions having an isolated essential singularity at $z = 0$ are

$$e^{1/z} = \sum_{n=0}^{\infty} \frac{1}{n!z^n} = 1 + \frac{1}{z} + \frac{1}{2!z^2} + \cdots$$

and

$$\sin\frac{1}{z} = \sum_{n=0}^{\infty}\frac{(-1)^n}{(2n+1)!z^{2n+1}} = \frac{1}{z} - \frac{1}{3!z^3} + \frac{1}{5!z^5} - + \cdots.$$

Section 16.1 provides further examples. In that section, Example 1 shows that $z^{-5}\sin z$ has a fourth-order pole at 0. Furthermore, Example 4 shows that $1/(z^3 - z^4)$ has a third-order pole at 0 and a Laurent series with infinitely many negative powers. This is no contradiction, since this series is valid for $|z| > 1$; it merely tells us that in classifying singularities it is quite important to consider the Laurent series valid ***in the immediate neighborhood*** of a singular point. In Example 4 this is the series (I), which has three negative powers. ■

The classification of singularities into poles and essential singularities is not merely a formal matter, because the behavior of an analytic function in a neighborhood of an essential singularity is entirely different from that in the neighborhood of a pole.

EXAMPLE 2 Behavior Near a Pole

$f(z) = 1/z^2$ has a pole at $z = 0$, and $|f(z)| \to \infty$ as $z \to 0$ in any manner. This illustrates the following theorem. ■

THEOREM 1

Poles

If $f(z)$ is analytic and has a pole at $z = z_0$, then $|f(z)| \to \infty$ as $z \to z_0$ in any manner.

The proof is left as an exercise (see Prob. 24).

EXAMPLE 3 Behavior Near an Essential Singularity

The function $f(z) = e^{1/z}$ has an essential singularity at $z = 0$. It has no limit for approach along the imaginary axis; it becomes infinite if $z \to 0$ through positive real values, but it approaches zero if $z \to 0$ through negative real values. It takes on any given value $c = c_0 e^{i\alpha} \neq 0$ in an arbitrarily small ϵ-neighborhood of $z = 0$. To see the latter, we set $z = re^{i\theta}$, and then obtain the following complex equation for r and θ, which we must solve:

$$e^{1/z} = e^{(\cos\theta - i\sin\theta)/r} = c_0 e^{i\alpha}.$$

Equating the absolute values and the arguments, we have $e^{(\cos\theta)/r} = c_0$, that is

$$\cos\theta = r\ln c_0, \qquad \text{and} \qquad -\sin\theta = \alpha r$$

respectively. From these two equations and $\cos^2\theta + \sin^2\theta = r^2(\ln c_0)^2 + \alpha^2 r^2 = 1$ we obtain the formulas

$$r^2 = \frac{1}{(\ln c_0)^2 + \alpha^2} \qquad \text{and} \qquad \tan\theta = -\frac{\alpha}{\ln c_0}.$$

Hence r can be made arbitrarily small by adding multiples of 2π to α, leaving c unaltered. This illustrates the very famous *Picard's theorem* (with $z = 0$ as the exceptional value). ■

THEOREM 2

Picard's Theorem

If $f(z)$ is analytic and has an isolated essential singularity at a point z_0, it takes on every value, with at most one exceptional value, in an arbitrarily small ϵ-neighborhood of z_0.

For the rather complicated proof, see Ref. [D4], vol. 2, p. 258. For historical information on Picard, see footnote 9 in Problem Set 1.7.

Removable Singularities. We say that a function $f(z)$ has a *removable singularity* at $z = z_0$ if $f(z)$ is not analytic at $z = z_0$, but can be made analytic there by assigning a suitable value $f(z_0)$. Such singularities are of no interest since they can be removed as just indicated. *Example:* $f(z) = (\sin z)/z$ becomes analytic at $z = 0$ if we define $f(0) = 1$.

Zeros of Analytic Functions

A **zero** of an analytic function $f(z)$ in a domain D is a $z = z_0$ in D such that $f(z_0) = 0$. A zero has **order** n if not only f but also the derivatives $f', f'', \cdots, f^{(n-1)}$ are all 0 at $z = z_0$ but $f^{(n)}(z_0) \neq 0$. A first-order zero is also called a **simple zero**. For a second-order zero, $f(z_0) = f'(z_0) = 0$ but $f''(z_0) \neq 0$. And so on.

EXAMPLE 4 Zeros

The function $1 + z^2$ has simple zeros at $\pm i$. The function $(1 - z^4)^2$ has second-order zeros at ± 1 and $\pm i$. The function $(z - a)^3$ has a third-order zero at $z = a$. The function e^z has no zeros (see Sec. 13.5). The function $\sin z$ has simple zeros at $0, \pm\pi, \pm 2\pi, \cdots$, and $\sin^2 z$ has second-order zeros at these points. The function $1 - \cos z$ has second-order zeros at $0, \pm 2\pi, \pm 4\pi, \cdots$, and the function $(1 - \cos z)^2$ has fourth-order zeros at these points. ■

Taylor Series at a Zero. At an nth-order zero $z = z_0$ of $f(z)$, the derivatives $f'(z_0), \cdots, f^{(n-1)}(z_0)$ are zero, by definition. Hence the first few coefficients $a_0, \cdots, a_{n-1}$ of the Taylor series (1), Sec. 15.4, are zero, too, whereas $a_n \neq 0$, so that this series takes the form

$$\begin{aligned} (3) \qquad f(z) &= a_n(z - z_0)^n + a_{n+1}(z - z_0)^{n+1} + \cdots \\ &= (z - z_0)^n \left[a_n + a_{n+1}(z - z_0) + a_{n+2}(z - z_0)^2 + \cdots\right] \qquad (a_n \neq 0). \end{aligned}$$

This is characteristic of such a zero, because, if $f(z)$ has such a Taylor series, it has an nth-order zero at $z = z_0$, as follows by differentiation.

Whereas nonisolated singularities may occur, for zeros we have

THEOREM 3 Zeros

The zeros of an analytic function $f(z)$ ($\not\equiv 0$) are isolated; that is, each of them has a neighborhood that contains no further zeros of $f(z)$.

PROOF The factor $(z - z_0)^n$ in (3) is zero only at $z = z_0$. The power series in the brackets $[\cdots]$ represents an analytic function (by Theorem 5 in Sec. 15.3), call it $g(z)$. Now $g(z_0) = a_n \neq 0$, since an analytic function is continuous, and because of this continuity, also $g(z) \neq 0$ in some neighborhood of $z = z_0$. Hence the same holds of $f(z)$. ■

This theorem is illustrated by the functions in Example 4.

Poles are often caused by zeros in the denominator. (*Example:* $\tan z$ has poles where $\cos z$ is zero.) This is a major reason for the importance of zeros. The key to the connection is the following theorem, whose proof follows from (3) (see Team Project 12).

THEOREM 4 Poles and Zeros

Let $f(z)$ be analytic at $z = z_0$ and have a zero of nth order at $z = z_0$. Then $1/f(z)$ has a pole of nth order at $z = z_0$; and so does $h(z)/f(z)$, provided $h(z)$ is analytic at $z = z_0$ and $h(z_0) \neq 0$.

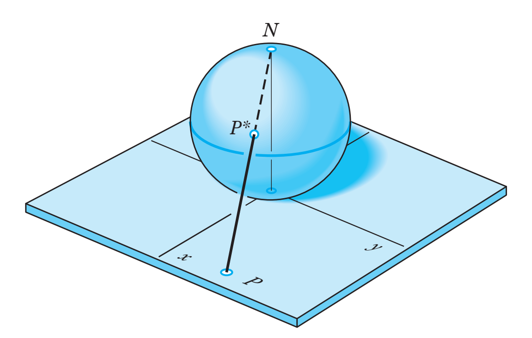


Fig. 372. Riemann sphere

Riemann Sphere. Point at Infinity

When we want to study complex functions for large $|z|$, the complex plane will generally become rather inconvenient. Then it may be better to use a representation of complex numbers on the so-called **Riemann sphere**. This is a sphere S of diameter 1 touching the complex z-plane at $z = 0$ (Fig. 372), and we let the image of a point P (a number z in the plane) be the intersection P^* of the segment PN with S, where N is the "North Pole" diametrically opposite to the origin in the plane. Then to each z there corresponds a point on S.

Conversely, each point on S represents a complex number z, except for N, which does not correspond to any point in the complex plane. This suggests that we introduce an additional point, called the **point at infinity** and denoted ∞ ("infinity") and let its image be N. The complex plane together with ∞ is called the **extended complex plane**. The complex plane is often called the *finite complex plane*, for distinction, or simply the *complex plane* as before. The sphere S is called the **Riemann sphere**. The mapping of the extended complex plane onto the sphere is known as a **stereographic projection**. (What is the image of the Northern Hemisphere? Of the Western Hemisphere? Of a straight line through the origin?)

Analytic or Singular at Infinity

If we want to investigate a function $f(z)$ for large $|z|$, we may now set $z = 1/w$ and investigate $f(z) = f(1/w) \equiv g(w)$ in a neighborhood of $w = 0$. We define $f(z)$ to be **analytic** or **singular at infinity** if $g(w)$ is analytic or singular, respectively, at $w = 0$. We also define

$$g(0) = \lim_{w \to 0} g(w) \tag{4}$$

if this limit exists.

Furthermore, we say that $f(z)$ has an *nth-order zero at infinity* if $f(1/w)$ has such a zero at $w = 0$. Similarly for poles and essential singularities.

EXAMPLE 5 **Functions Analytic or Singular at Infinity. Entire and Meromorphic Functions**

The function $f(z) = 1/z^2$ is analytic at ∞ since $g(w) = f(1/w) = w^2$ is analytic at $w = 0$, and $f(z)$ has a second-order zero at ∞. The function $f(z) = z^3$ is singular at ∞ and has a third-order pole there since the function $g(w) = f(1/w) = 1/w^3$ has such a pole at $w = 0$. The function e^z has an essential singularity at ∞ since $e^{1/w}$ has such a singularity at $w = 0$. Similarly, $\cos z$ and $\sin z$ have an essential singularity at ∞.

Recall that an **entire function** is one that is analytic everywhere in the (finite) complex plane. Liouville's theorem (Sec. 14.4) tells us that the only *bounded* entire functions are the constants, hence any nonconstant entire function must be unbounded. Hence it has a singularity at ∞, a pole if it is a polynomial or an essential singularity if it is not. The functions just considered are typical in this respect.

An analytic function whose only singularities in the finite plane are poles is called a **meromorphic function.** Examples are rational functions with nonconstant denominator, $\tan z$, $\cot z$, $\sec z$, and $\csc z$. ■

In this section we used Laurent series for investigating singularities. In the next section we shall use these series for an elegant integration method.

PROBLEM SET 16.2

1–10 ZEROS

Determine the location and order of the zeros.

1. $\sin^4 \frac{1}{2}z$
2. $(z^4 - 81)^3$
3. $(z + 81i)^4$
4. $\tan^2 2z$
5. $z^{-2} \sin^2 \pi z$
6. $\cosh^4 z$
7. $z^4 + (1 - 8i)z^2 - 8i$
8. $(\sin z - 1)^3$
9. $\sin 2z \cos 2z$
10. $(z^2 - 8)^3(\exp(z^2) - 1)$
11. **Zeros.** If $f(z)$ is analytic and has a zero of order n at $z = z_0$, show that $f^2(z)$ has a zero of order $2n$ at z_0.
12. **TEAM PROJECT. Zeros. (a) Derivative.** Show that if $f(z)$ has a zero of order $n > 1$ at $z = z_0$, then $f'(z)$ has a zero of order $n - 1$ at z_0.

    (b) Poles and zeros. Prove Theorem 4.

    (c) Isolated k-points. Show that the points at which a nonconstant analytic function $f(z)$ has a given value k are isolated.

    (d) Identical functions. If $f_1(z)$ and $f_2(z)$ are analytic in a domain D and equal at a sequence of points z_n in D that converges in D, show that $f_1(z) \equiv f_2(z)$ in D.

13–22 SINGULARITIES

Determine the location of the singularities, including those at infinity. For poles also state the order. Give reasons.

13. $\dfrac{1}{(z + 2i)^2} - \dfrac{z}{z - i} + \dfrac{z + 1}{(z - i)^2}$
14. $e^{z-i} + \dfrac{2}{z - i} - \dfrac{8}{(z - i)^3}$
15. $z \exp(1/(z - 1 - i)^2)$
16. $\tan \pi z$
17. $\cot^4 z$
18. $z^3 \exp\left(\dfrac{1}{z - 1}\right)$
19. $1/(e^z - e^{2z})$
20. $1/(\cos z - \sin z)$
21. $e^{1/(z-1)}/(e^z - 1)$
22. $(z - \pi)^{-1} \sin z$
23. **Essential singularity.** Discuss e^{1/z^2} in a similar way as $e^{1/z}$ is discussed in Example 3 of the text.
24. **Poles.** Verify Theorem 1 for $f(z) = z^{-3} - z^{-1}$. Prove Theorem 1.
25. **Riemann sphere.** Assuming that we let the image of the x-axis be the meridians 0° and 180°, describe and sketch (or graph) the images of the following regions on the Riemann sphere: **(a)** $|z| > 100$, **(b)** the lower half-plane, **(c)** $\frac{1}{2} \leqq |z| \leqq 2$.

16.3 Residue Integration Method

We now cover a second method of evaluating complex integrals. Recall that we solved complex integrals directly by Cauchy's integral formula in Sec. 14.3. In Chapter 15 we learned about power series and especially Taylor series. We generalized Taylor series to Laurent series (Sec. 16.1) and investigated singularities and zeroes of various functions (Sec. 16.2). Our hard work has paid off and we see how much of the theoretical groundwork comes together in evaluating complex integrals by the residue method.

The purpose of Cauchy's residue integration method is the evaluation of integrals

$$\oint_C f(z)\, dz$$

taken around a simple closed path C. The idea is as follows.

If $f(z)$ is analytic everywhere on C and inside C, such an integral is zero by Cauchy's integral theorem (Sec. 14.2), and we are done.

The situation changes if $f(z)$ has a singularity at a point $z = z_0$ inside C but is otherwise analytic on C and inside C as before. Then $f(z)$ has a Laurent series

$$f(z) = \sum_{n=0}^{\infty} a_n(z - z_0)^n + \frac{b_1}{z - z_0} + \frac{b_2}{(z - z_0)^2} + \cdots$$

that converges for all points near $z = z_0$ (except at $z = z_0$ itself), in some domain of the form $0 < |z - z_0| < R$ (sometimes called a **deleted neighborhood**, an old-fashioned term that we shall not use). Now comes the key idea. The coefficient b_1 of the first negative power $1/(z - z_0)$ of this Laurent series is given by the integral formula (2) in Sec. 16.1 with $n = 1$, namely,

$$b_1 = \frac{1}{2\pi i} \oint_C f(z)\, dz.$$

Now, since we can obtain Laurent series by various methods, without using the integral formulas for the coefficients (see the examples in Sec. 16.1), we can find b_1 by one of those methods and then use the formula for b_1 for evaluating the integral, that is,

(1) $$\oint_C f(z)\, dz = 2\pi i b_1.$$

Here we integrate counterclockwise around a simple closed path C that contains $z = z_0$ in its interior (but no other singular points of $f(z)$ on or inside C!).

The coefficient b_1 is called the **residue** of $f(z)$ at $z = z_0$ and we denote it by

(2) $$b_1 = \operatorname*{Res}_{z=z_0} f(z).$$

EXAMPLE 1 Evaluation of an Integral by Means of a Residue

Integrate the function $f(z) = z^{-4} \sin z$ counterclockwise around the unit circle C.

Solution. From (14) in Sec. 15.4 we obtain the Laurent series

$$f(z) = \frac{\sin z}{z^4} = \frac{1}{z^3} - \frac{1}{3!z} + \frac{1}{5!} - \frac{z^3}{7!} + - \cdots$$

which converges for $|z| > 0$ (that is, for all $z \neq 0$). This series shows that $f(z)$ has a pole of third order at $z = 0$ and the residue $b_1 = -\frac{1}{3}!$. From (1) we thus obtain the answer

$$\oint_C \frac{\sin z}{z^4}\, dz = 2\pi i b_1 = -\frac{\pi i}{3}.$$ ■

EXAMPLE 2 CAUTION! Use the Right Laurent Series!

Integrate $f(z) = 1/(z^3 - z^4)$ clockwise around the circle C: $|z| = \frac{1}{2}$.

Solution. $z^3 - z^4 = z^3(1 - z)$ shows that $f(z)$ is singular at $z = 0$ and $z = 1$. Now $z = 1$ lies outside C. Hence it is of no interest here. So we need the residue of $f(z)$ at 0. We find it from the Laurent series that converges for $0 < |z| < 1$. This is series (I) in Example 4, Sec. 16.1,

$$\frac{1}{z^3 - z^4} = \frac{1}{z^3} + \frac{1}{z^2} + \frac{1}{z} + 1 + z + \cdots \qquad (0 < |z| < 1).$$

We see from it that this residue is 1. Clockwise integration thus yields

$$\oint_C \frac{dz}{z^3 - z^4} = -2\pi i \operatorname*{Res}_{z=0} f(z) = -2\pi i.$$

CAUTION! Had we used the wrong series (II) in Example 4, Sec. 16.1,

$$\frac{1}{z^3 - z^4} = -\frac{1}{z^4} - \frac{1}{z^5} - \frac{1}{z^6} - \cdots \qquad (|z| > 1),$$

we would have obtained the wrong answer, 0, because this series has no power $1/z$. ■

Formulas for Residues

To calculate a residue at a pole, we need not produce a whole Laurent series, but, more economically, we can derive formulas for residues once and for all.

Simple Poles at z_0. A first formula for the residue at a simple pole is

(3)
$$\operatorname*{Res}_{z=z_0} f(z) = b_1 = \lim_{z \to z_0} (z - z_0) f(z). \qquad \text{(Proof below).}$$

A second formula for the residue at a simple pole is

(4)
$$\operatorname*{Res}_{z=z_0} f(z) = \operatorname*{Res}_{z=z_0} \frac{p(z)}{q(z)} = \frac{p(z_0)}{q'(z_0)}. \qquad \text{(Proof below).}$$

In (4) we assume that $f(z) = p(z)/q(z)$ with $p(z_0) \neq 0$ and $q(z)$ has a simple zero at z_0, so that $f(z)$ has a simple pole at z_0 by Theorem 4 in Sec. 16.2.

PROOF We prove (3). For a simple pole at $z = z_0$ the Laurent series (1), Sec. 16.1, is

$$f(z) = \frac{b_1}{z - z_0} + a_0 + a_1(z - z_0) + a_2(z - z_0)^2 + \cdots \qquad (0 < |z - z_0| < R).$$

Here $b_1 \neq 0$. (Why?) Multiplying both sides by $z - z_0$ and then letting $z \to z_0$, we obtain the formula (3):

$$\lim_{z \to z_0} (z - z_0) f(z) = b_1 + \lim_{z \to z_0} (z - z_0)[a_0 + a_1(z - z_0) + \cdots] = b_1$$

where the last equality follows from continuity (Theorem 1, Sec. 15.3).

We prove (4). The Taylor series of $q(z)$ at a simple zero z_0 is

$$q(z) = (z - z_0) q'(z_0) + \frac{(z - z_0)^2}{2!} q''(z_0) + \cdots.$$

Substituting this into $f = p/q$ and then f into (3) gives

$$\operatorname*{Res}_{z=z_0} f(z) = \lim_{z \to z_0} (z - z_0) \frac{p(z)}{q(z)} = \lim_{z \to z_0} \frac{(z - z_0) p(z)}{(z - z_0)[q'(z_0) + (z - z_0) q''(z_0)/2 + \cdots]}.$$

$z - z_0$ cancels. By continuity, the limit of the denominator is $q'(z_0)$ and (4) follows. ■

EXAMPLE 3 **Residue at a Simple Pole**

$f(z) = (9z + i)/(z^3 + z)$ has a simple pole at i because $z^2 + 1 = (z + i)(z - i)$, and (3) gives the residue

$$\operatorname*{Res}_{z=i} \frac{9z + i}{z(z^2 + 1)} = \lim_{z \to i} (z - i) \frac{9z + i}{z(z + i)(z - i)} = \left[\frac{9z + i}{z(z + i)}\right]_{z=i} = \frac{10i}{-2} = -5i.$$

By (4) with $p(i) = 9i + i$ and $q'(z) = 3z^2 + 1$ we confirm the result,

$$\operatorname*{Res}_{z=i} \frac{9z + i}{z(z^2 + 1)} = \left[\frac{9z + i}{3z^2 + 1}\right]_{z=i} = \frac{10i}{-2} = -5i.$$ ■

Poles of Any Order at z_0. The residue of $f(z)$ at an mth-order pole at z_0 is

(5) $$\operatorname*{Res}_{z=z_0} f(z) = \frac{1}{(m - 1)!} \lim_{z \to z_0} \left\{ \frac{d^{m-1}}{dz^{m-1}} \left[(z - z_0)^m f(z) \right] \right\}.$$

In particular, for a second-order pole $(m = 2)$,

(5*) $$\operatorname*{Res}_{z=z_0} f(z) = \lim_{z \to z_0} \{[(z - z_0)^2 f(z)]'\}.$$

PROOF We prove (5). The Laurent series of $f(z)$ converging near z_0 (except at z_0 itself) is (Sec. 16.2)

$$f(z) = \frac{b_m}{(z - z_0)^m} + \frac{b_{m-1}}{(z - z_0)^{m-1}} + \cdots + \frac{b_1}{z - z_0} + a_0 + a_1(z - z_0) + \cdots$$

where $b_m \neq 0$. The residue wanted is b_1. Multiplying both sides by $(z - z_0)^m$ gives

$$(z - z_0)^m f(z) = b_m + b_{m-1}(z - z_0) + \cdots + b_1(z - z_0)^{m-1} + a_0(z - z_0)^m + \cdots.$$

We see that b_1 is now the coefficient of the power $(z - z_0)^{m-1}$of the power series of $g(z) = (z - z_0)^m f(z)$. Hence Taylor's theorem (Sec. 15.4) gives (5):

$$\begin{aligned} b_1 &= \frac{1}{(m - 1)!} g^{(m-1)}(z_0) \\ &= \frac{1}{(m - 1)!} \frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)]. \end{aligned}$$ ■

EXAMPLE 4 **Residue at a Pole of Higher Order**

$f(z) = 50z/(z^3 + 2z^2 - 7z + 4)$ has a pole of second order at $z = 1$ because the denominator equals $(z + 4)(z - 1)^2$ (verify!). From (5*) we obtain the residue

$$\operatorname*{Res}_{z=1} f(z) = \lim_{z \to 1} \frac{d}{dz}[(z - 1)^2 f(z)] = \lim_{z \to 1} \frac{d}{dz}\left(\frac{50z}{z + 4}\right) = \frac{200}{5^2} = 8.$$ ■

Several Singularities Inside the Contour. Residue Theorem

Residue integration can be extended from the case of a single singularity to the case of several singularities within the contour C. This is the purpose of the residue theorem. The extension is surprisingly simple.

THEOREM 1

Residue Theorem

Let $f(z)$ be analytic inside a simple closed path C and on C, except for finitely many singular points $z_1, z_2, \cdots, z_k$ inside C. Then the integral of $f(z)$ taken counterclockwise around C equals $2\pi i$ times the sum of the residues of $f(z)$ at $z_1, \cdots, z_k$:

$$\oint_C f(z)\,dz = 2\pi i \sum_{j=1}^{k} \operatorname*{Res}_{z=z_j} f(z). \tag{6}$$

PROOF We enclose each of the singular points z_j in a circle C_j with radius small enough that those k circles and C are all separated (Fig. 373 where $k = 3$). Then $f(z)$ is analytic in the multiply connected domain D bounded by C and $C_1, \cdots, C_k$ and on the entire boundary of D. From Cauchy's integral theorem we thus have

$$\oint_C f(z)\,dz + \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz + \cdots + \oint_{C_k} f(z)\,dz = 0, \tag{7}$$

the integral along C being taken *counterclockwise* and the other integrals *clockwise* (as in Figs. 354 and 355, Sec. 14.2). We take the integrals over $C_1, \cdots, C_k$ to the right and compensate the resulting minus sign by reversing the sense of integration. Thus,

$$\oint_C f(z)\,dz = \oint_{C_1} f(z)\,dz + \oint_{C_2} f(z)\,dz + \cdots + \oint_{C_k} f(z)\,dz \tag{8}$$

where all the integrals are now taken counterclockwise. By (1) and (2),

$$\oint_{C_j} f(z)\,dz = 2\pi i \operatorname*{Res}_{z=z_j} f(z), \qquad j = 1, \cdots, k,$$

so that (8) gives (6) and the residue theorem is proved. ■

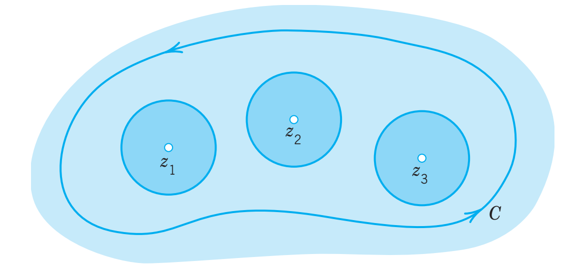


Fig. 373. Residue theorem

This important theorem has various applications in connection with complex and real integrals. Let us first consider some complex integrals. (Real integrals follow in the next section.)

EXAMPLE 5 **Integration by the Residue Theorem. Several Contours**

Evaluate the following integral counterclockwise around any simple closed path such that (a) 0 and 1 are inside C, (b) 0 is inside, 1 outside, (c) 1 is inside, 0 outside, (d) 0 and 1 are outside.

$$\oint_C \frac{4 - 3z}{z^2 - z}\, dz$$

Solution. The integrand has simple poles at 0 and 1, with residues [by (3)]

$$\operatorname*{Res}_{z=0} \frac{4 - 3z}{z(z - 1)} = \left[\frac{4 - 3z}{z - 1}\right]_{z=0} = -4, \qquad \operatorname*{Res}_{z=1} \frac{4 - 3z}{z(z - 1)} = \left[\frac{4 - 3z}{z}\right]_{z=1} = 1.$$

[Confirm this by (4).] *Answer:* (a) $2\pi i(-4 + 1) = -6\pi i$, (b) $-8\pi i$, (c) $2\pi i$, (d) 0. ■

EXAMPLE 6 **Another Application of the Residue Theorem**

Integrate $(\tan z)/(z^2 - 1)$ counterclockwise around the circle C: $|z| = \frac{3}{2}$.

Solution. $\tan z$ is not analytic at $\pm\pi/2, \pm 3\pi/2, \cdots$, but all these points lie outside the contour C. Because of the denominator $z^2 - 1 = (z - 1)(z + 1)$ the given function has simple poles at ± 1. We thus obtain from (4) and the residue theorem

$$\oint_C \frac{\tan z}{z^2 - 1}\, dz = 2\pi i \left(\operatorname*{Res}_{z=1} \frac{\tan z}{z^2 - 1} + \operatorname*{Res}_{z=-1} \frac{\tan z}{z^2 - 1}\right)$$

$$= 2\pi i \left(\frac{\tan z}{2z}\bigg|_{z=1} + \frac{\tan z}{2z}\bigg|_{z=-1}\right)$$

$$= 2\pi i \tan 1 = 9.7855i.$$ ■

EXAMPLE 7 **Poles and Essential Singularities**

Evaluate the following integral, where C is the ellipse $9x^2 + y^2 = 9$ (counterclockwise, sketch it).

$$\oint_C \left(\frac{ze^{\pi z}}{z^4 - 16} + ze^{\pi/z}\right) dz.$$

Solution. Since $z^4 - 16 = 0$ at $\pm 2i$ and ± 2, the first term of the integrand has simple poles at $\pm 2i$ inside C, with residues [by (4); note that $e^{2\pi i} = 1$]

$$\operatorname*{Res}_{z=2i} \frac{ze^{\pi z}}{z^4 - 16} = \left[\frac{ze^{\pi z}}{4z^3}\right]_{z=2i} = -\frac{1}{16},$$

$$\operatorname*{Res}_{z=-2i} \frac{ze^{\pi z}}{z^4 - 16} = \left[\frac{ze^{\pi z}}{4z^3}\right]_{z=-2i} = -\frac{1}{16}$$

and simple poles at ± 2, which lie outside C, so that they are of no interest here. The second term of the integrand has an essential singularity at 0, with residue $\pi^2/2$ as obtained from

$$ze^{\pi/z} = z\left(1 + \frac{\pi}{z} + \frac{\pi^2}{2!z^2} + \frac{\pi^3}{3!z^3} + \cdots\right) = z + \pi + \frac{\pi^2}{2} \cdot \frac{1}{z} + \cdots \qquad (|z| > 0).$$

Answer: $2\pi i(-\frac{1}{16} - \frac{1}{16} + \frac{1}{2}\pi^2) = \pi(\pi^2 - \frac{1}{4})i = 30.221i$ by the residue theorem. ■

PROBLEM SET 16.3

1. Verify the calculations in Example 3 and find the other residues.
2. Verify the calculations in Example 4 and find the other residue.

3–12 RESIDUES

Find all the singularities in the finite plane and the corresponding residues. Show the details.

3. $\frac{\sin 2z}{z^6}$
4. $\frac{\cos z}{z^4}$
5. $\frac{8}{1 + z^2}$
6. $\tan z$
7. $\cot \pi z$
8. $\frac{\pi}{(z^2 - 1)^2}$
9. $\frac{1}{1 - e^z}$
10. $\frac{z^4}{z^2 - iz + 2}$
11. $\frac{e^z}{(z - \pi i)^3}$
12. $e^{1/(1-z)}$
13. **CAS PROJECT. Residue at a Pole.** Write a program for calculating the residue at a pole of any order in the finite plane. Use it for solving Probs. 5–10.

14–25 RESIDUE INTEGRATION

Evaluate (counterclockwise). Show the details.

14. $\oint_C \frac{z - 23}{z^2 - 4z - 5}\, dz, \quad C: |z - 2 - i| = 3.2$
15. $\oint_C \tan 2\pi z\, dz, \quad C: |z - 0.2| = 0.2$
16. $\oint_C e^{1/z}\, dz, \quad C$: the unit circle
17. $\oint_C \frac{e^z}{\cos z}\, dz, \quad C: |z - \pi i/2| = 4.5$
18. $\oint_C \frac{z + 1}{z^4 - 2z^3}\, dz, \quad C: |z - 1| = 2$
19. $\oint_C \frac{\sinh z}{2z - i}\, dz, \quad C: |z - 2i| = 2$
20. $\oint_C \frac{dz}{(z^2 + 1)^3}, \quad C: |z - i| = 3$
21. $\oint_C \frac{\cos \pi z}{z^5}\, dz, \quad C: |z| = \frac{1}{2}$
22. $\oint_C \frac{z^2 \sin z}{4z^2 - 1}\, dz, \quad C$ the unit circle
23. $\oint_C \frac{30z^2 - 23z + 5}{(2z - 1)^2(3z - 1)}\, dz, \quad C$ the unit circle
24. $\oint_C \frac{\exp(-z^2)}{\sin 4z}\, dz, \quad C: |z| = 1.5$
25. $\oint_C \frac{z \cosh \pi z}{z^4 + 13z^2 + 36}\, dz, \quad |z| = \pi$

16.4 Residue Integration of Real Integrals

Surprisingly, residue integration can also be used to evaluate certain classes of complicated real integrals. This shows an advantage of complex analysis over real analysis or calculus.

Integrals of Rational Functions of $\cos\theta$ and $\sin\theta$

We first consider integrals of the type

$$J = \int_0^{2\pi} F(\cos\theta, \sin\theta)\, d\theta \qquad (1)$$

where $F(\cos\theta, \sin\theta)$ is a real rational function of $\cos\theta$ and $\sin\theta$ [for example, $(\sin^2\theta)/(5 - 4\cos\theta)$] and is finite (does not become infinite) on the interval of integration. Setting $e^{i\theta} = z$, we obtain

(2)
$$\cos\theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}) = \frac{1}{2}\left(z + \frac{1}{z}\right)$$
$$\sin\theta = \frac{1}{2i}(e^{i\theta} - e^{-i\theta}) = \frac{1}{2i}\left(z - \frac{1}{z}\right).$$

Since F is rational in $\cos\theta$ and $\sin\theta$, Eq. (2) shows that F is now a rational function of z, say, $f(z)$. Since $dz/d\theta = ie^{i\theta}$, we have $d\theta = dz/iz$ and the given integral takes the form

(3)
$$J = \oint_C f(z)\,\frac{dz}{iz}$$

and, as θ ranges from 0 to 2π in (1), the variable $z = e^{i\theta}$ ranges counterclockwise once around the unit circle $|z| = 1$. (Review Sec. 13.5 if necessary.)

EXAMPLE 1 An Integral of the Type (1)

Show by the present method that $\displaystyle\int_0^{2\pi} \frac{d\theta}{\sqrt{2} - \cos\theta} = 2\pi.$

Solution. We use $\cos\theta = \frac{1}{2}(z + 1/z)$ and $d\theta = dz/iz$. Then the integral becomes

$$\oint_C \frac{dz/iz}{\sqrt{2} - \dfrac{1}{2}\left(z + \dfrac{1}{z}\right)} = \oint_C \frac{dz}{-\dfrac{i}{2}(z^2 - 2\sqrt{2}z + 1)}$$
$$= -\frac{2}{i}\oint_C \frac{dz}{(z - \sqrt{2} - 1)(z - \sqrt{2} + 1)}.$$

We see that the integrand has a simple pole at $z_1 = \sqrt{2} + 1$ outside the unit circle C, so that it is of no interest here, and another simple pole at $z_2 = \sqrt{2} - 1$ (where $z - \sqrt{2} + 1 = 0$) inside C with residue [by (3), Sec. 16.3]

$$\operatorname*{Res}_{z=z_2} \frac{1}{(z - \sqrt{2} - 1)(z - \sqrt{2} + 1)} = \left[\frac{1}{z - \sqrt{2} - 1}\right]_{z=\sqrt{2}-1}$$
$$= -\frac{1}{2}.$$

Answer: $2\pi i(-2/i)(-\frac{1}{2}) = 2\pi$. (Here $-2/i$ is the factor in front of the last integral.) ■

As another large class, let us consider real integrals of the form

(4)
$$\int_{-\infty}^{\infty} f(x)\,dx.$$

Such an integral, whose interval of integration is not finite is called an **improper integral**, and it has the meaning

(5′)
$$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{a\to-\infty}\int_a^0 f(x)\,dx + \lim_{b\to\infty}\int_0^b f(x)\,dx.$$

If both limits exist, we may couple the two independent passages to $-\infty$ and ∞, and write

(5) $$\int_{-\infty}^{\infty} f(x)\,dx = \lim_{R\to\infty} \int_{-R}^{R} f(x)\,dx.$$

The limit in (5) is called the **Cauchy principal value** of the integral. It is written

$$\text{pr. v.} \int_{-\infty}^{\infty} f(x)\,dx.$$

It may exist even if the limits in (5′) do not. *Example:*

$$\lim_{R\to\infty} \int_{-R}^{R} x\,dx = \lim_{R\to\infty} \left(\frac{R^2}{2} - \frac{R^2}{2}\right) = 0, \qquad \text{but} \qquad \lim_{b\to\infty} \int_{0}^{b} x\,dx = \infty.$$

We assume that the function $f(x)$ in (4) is a real rational function whose denominator is different from zero for all real x and is of degree at least two units higher than the degree of the numerator. Then the limits in (5′) exist, and we may start from (5). We consider the corresponding contour integral

(5*) $$\oint_C f(z)\,dz$$

around a path C in Fig. 374. Since $f(x)$ is rational, $f(z)$ has finitely many poles in the upper half-plane, and if we choose R large enough, then C encloses all these poles. By the residue theorem we then obtain

$$\oint_C f(z)\,dz = \int_S f(z)\,dz + \int_{-R}^{R} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z)$$

where the sum consists of all the residues of $f(z)$ at the points in the upper half-plane at which $f(z)$ has a pole. From this we have

(6) $$\int_{-R}^{R} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z) - \int_S f(z)\,dz.$$

We prove that, if $R \to \infty$, the value of the integral over the semicircle S approaches zero. If we set $z = Re^{i\theta}$, then S is represented by $R = \text{const}$, and as z ranges along S, the variable θ ranges from 0 to π. Since, by assumption, the degree of the denominator of $f(z)$ is at least two units higher than the degree of the numerator, we have

$$|f(z)| < \frac{k}{|z|^2} \qquad (|z| = R > R_0)$$

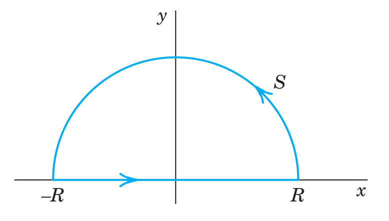


Fig. 374. Path C of the contour integral in (5*)

for sufficiently large constants k and R_0. By the ML-inequality in Sec. 14.1,

$$\left| \int_S f(z)\, dz \right| < \frac{k}{R^2} \pi R = \frac{k\pi}{R} \qquad (R > R_0).$$

Hence, as R approaches infinity, the value of the integral over S approaches zero, and (5) and (6) yield the result

(7)
$$\int_{-\infty}^{\infty} f(x)\, dx = 2\pi i \sum \operatorname{Res} f(z)$$

where we sum over all the residues of $f(z)$ at the poles of $f(z)$ in the upper half-plane.

EXAMPLE 2 An Improper Integral from 0 to ∞

Using (7), show that

$$\int_0^{\infty} \frac{dx}{1 + x^4} = \frac{\pi}{2\sqrt{2}}.$$

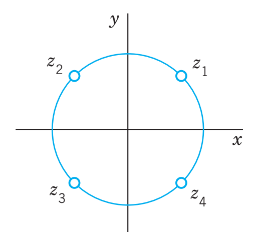


Fig. 375. Example 2

Solution. Indeed, $f(z) = 1/(1 + z^4)$ has four simple poles at the points (make a sketch)

$$z_1 = e^{\pi i/4}, \qquad z_2 = e^{3\pi i/4}, \qquad z_3 = e^{-3\pi i/4}, \qquad z_4 = e^{-\pi i/4}.$$

The first two of these poles lie in the upper half-plane (Fig. 375). From (4) in the last section we find the residues

$$\operatorname*{Res}_{z=z_1} f(z) = \left[\frac{1}{(1 + z^4)'} \right]_{z=z_1} = \left[\frac{1}{4z^3} \right]_{z=z_1} = \frac{1}{4} e^{-3\pi i/4} = -\frac{1}{4} e^{\pi i/4}.$$

$$\operatorname*{Res}_{z=z_2} f(z) = \left[\frac{1}{(1 + z^4)'} \right]_{z=z_2} = \left[\frac{1}{4z^3} \right]_{z=z_2} = \frac{1}{4} e^{-9\pi i/4} = \frac{1}{4} e^{-\pi i/4}.$$

(Here we used $e^{\pi i} = -1$ and $e^{-2\pi i} = 1$.) By (1) in Sec. 13.6 and (7) in this section,

$$\int_{-\infty}^{\infty} \frac{dx}{1 + x^4} = -\frac{2\pi i}{4} (e^{\pi i/4} - e^{-\pi i/4}) = -\frac{2\pi i}{4} \cdot 2i \sin \frac{\pi}{4} = \pi \sin \frac{\pi}{4} = \frac{\pi}{\sqrt{2}}.$$

Since $1/(1 + x^4)$ is an even function, we thus obtain, as asserted,

$$\int_0^\infty \frac{dx}{1+x^4} = \frac{1}{2}\int_{-\infty}^\infty \frac{dx}{1+x^4} = \frac{\pi}{2\sqrt{2}}.$$ ■

Fourier Integrals

The method of evaluating (4) by creating a closed contour (Fig. 374) and "blowing it up" extends to integrals

$$\text{(8)} \qquad \int_{-\infty}^{\infty} f(x)\cos sx\, dx \quad \text{and} \quad \int_{-\infty}^{\infty} f(x)\sin sx\, dx \qquad (s \text{ real})$$

as they occur in connection with the Fourier integral (Sec. 11.7).

If $f(x)$ is a rational function satisfying the assumption on the degree as for (4), we may consider the corresponding integral

$$\oint_C f(z)e^{isz}\, dz \qquad (s \text{ real and positive})$$

over the contour C in Fig. 374. Instead of (7) we now get

$$\text{(9)} \qquad \int_{-\infty}^{\infty} f(x)e^{isx}\, dx = 2\pi i \sum \operatorname{Res}\,[f(z)e^{isz}] \qquad (s > 0)$$

where we sum the residues of $f(z)e^{isz}$ at its poles in the upper half-plane. Equating the real and the imaginary parts on both sides of (9), we have

$$\textbf{(10)} \qquad \begin{aligned} \int_{-\infty}^{\infty} f(x)\cos sx\, dx &= -2\pi \sum \operatorname{Im}\operatorname{Res}\,[f(z)e^{isz}], \\ \int_{-\infty}^{\infty} f(x)\sin sx\, dx &= 2\pi \sum \operatorname{Re}\operatorname{Res}\,[f(z)e^{isz}]. \end{aligned} \qquad (s > 0)$$

To establish (9), we must show [as for (4)] that the value of the integral over the semicircle S in Fig. 374 approaches 0 as $R \to \infty$. Now $s > 0$ and S lies in the upper half-plane $y \geqq 0$. Hence

$$|e^{isz}| = |e^{is(x+iy)}| = |e^{isx}||e^{-sy}| = 1 \cdot e^{-sy} \leqq 1 \qquad (s > 0, \quad y \geqq 0).$$

From this we obtain the inequality $|f(z)e^{isz}| = |f(z)||e^{isz}| \leqq |f(z)| \quad (s > 0, \quad y \geqq 0)$. This reduces our present problem to that for (4). Continuing as before gives (9) and (10). ■

EXAMPLE 3 An Application of (10)

Show that

$$\int_{-\infty}^{\infty} \frac{\cos sx}{k^2 + x^2}\, dx = \frac{\pi}{k} e^{-ks}, \qquad \int_{-\infty}^{\infty} \frac{\sin sx}{k^2 + x^2}\, dx = 0 \qquad (s > 0, k > 0).$$

Solution. In fact, $e^{isz}/(k^2 + z^2)$ has only one pole in the upper half-plane, namely, a simple pole at $z = ik$, and from (4) in Sec. 16.3 we obtain

$$\operatorname*{Res}_{z=ik} \frac{e^{isz}}{k^2 + z^2} = \left[\frac{e^{isz}}{2z}\right]_{z=ik} = \frac{e^{-ks}}{2ik}.$$

Thus

$$\int_{-\infty}^{\infty} \frac{e^{isx}}{k^2 + x^2}\, dx = 2\pi i \frac{e^{-ks}}{2ik} = \frac{\pi}{k} e^{-ks}.$$

Since $e^{isx} = \cos sx + i \sin sx$, this yields the above results [see also (15) in Sec. 11.7.] ■

Another Kind of Improper Integral

We consider an improper integral

$$\int_A^B f(x)\, dx \tag{11}$$

whose integrand becomes infinite at a point a in the interval of integration,

$$\lim_{x \to a} |f(x)| = \infty.$$

By definition, this integral (11) means

$$\int_A^B f(x)\, dx = \lim_{\epsilon \to 0} \int_A^{a-\epsilon} f(x)\, dx + \lim_{\eta \to 0} \int_{a+\eta}^B f(x)\, dx \tag{12}$$

where both ϵ and η approach zero independently and through positive values. It may happen that neither of these two limits exists if ϵ and η go to 0 independently, but the limit

$$\lim_{\epsilon \to 0} \left[\int_A^{a-\epsilon} f(x)\, dx + \int_{a+\epsilon}^B f(x)\, dx\right] \tag{13}$$

exists. This is called the **Cauchy principal value** of the integral. It is written

$$\text{pr. v.} \int_A^B f(x)\, dx.$$

For example,

$$\text{pr. v.} \int_{-1}^{1} \frac{dx}{x^3} = \lim_{\epsilon \to 0} \left[\int_{-1}^{-\epsilon} \frac{dx}{x^3} + \int_{\epsilon}^{1} \frac{dx}{x^3}\right] = 0;$$

the principal value exists, although the integral itself has no meaning.

In the case of simple poles on the real axis we shall obtain a formula for the principal value of an integral from $-\infty$ to ∞. This formula will result from the following theorem.

THEOREM 1

Simple Poles on the Real Axis

If $f(z)$ *has a simple pole at* $z = a$ *on the real axis, then* (Fig. 376)

$$\lim_{r \to 0} \int_{C_2} f(z)\, dz = \pi i \operatorname*{Res}_{z=a} f(z).$$

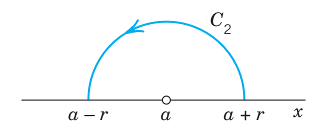


Fig. 376. Theorem 1

PROOF By the definition of a simple pole (Sec. 16.2) the integrand $f(z)$ has for $0 < |z - a| < R$ the Laurent series

$$f(z) = \frac{b_1}{z - a} + g(z), \qquad b_1 = \operatorname*{Res}_{z=a} f(z).$$

Here $g(z)$ is analytic on the semicircle of integration (Fig. 376)

$$C_2: \quad z = a + re^{i\theta}, \qquad 0 \leqq \theta \leqq \pi$$

and for all z between C_2 and the x-axis, and thus bounded on C_2, say, $|g(z)| \leqq M$. By integration,

$$\int_{C_2} f(z)\, dz = \int_0^{\pi} \frac{b_1}{re^{i\theta}}\, ire^{i\theta}\, d\theta + \int_{C_2} g(z)\, dz = b_1 \pi i + \int_{C_2} g(z)\, dz.$$

The second integral on the right cannot exceed $M\pi r$ in absolute value, by the ML-inequality (Sec. 14.1), and $ML = M\pi r \to 0$ as $r \to 0$. ■

Figure 377 shows the idea of applying Theorem 1 to obtain the principal value of the integral of a rational function $f(x)$ from $-\infty$ to ∞. For sufficiently large R the integral over the entire contour in Fig. 377 has the value J given by $2\pi i$ times the sum of the residues of $f(z)$ at the singularities in the upper half-plane. We assume that $f(x)$ satisfies the degree condition imposed in connection with (4). Then the value of the integral over the large

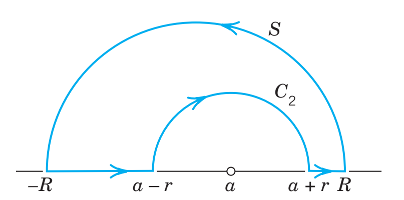


Fig. 377. Application of Theorem 1

semicircle S approaches 0 as $R \to \infty$. For $r \to 0$ the integral over C_2 (clockwise!) approaches the value

$$K = -\pi i \operatorname*{Res}_{z=a} f(z)$$

by Theorem 1. Together this shows that the principal value P of the integral from $-\infty$ to ∞ plus K equals J; hence $P = J - K = J + \pi i \operatorname{Res}_{z=a} f(z)$. If $f(z)$ has several simple poles on the real axis, then K will be $-\pi i$ times the sum of the corresponding residues. Hence the desired formula is

$$\text{(14)} \qquad \text{pr. v.} \int_{-\infty}^{\infty} f(x)\,dx = 2\pi i \sum \operatorname{Res} f(z) + \pi i \sum \operatorname{Res} f(z)$$

where the first sum extends over all poles in the upper half-plane and the second over all poles on the real axis, the latter being simple by assumption.

EXAMPLE 4 Poles on the Real Axis

Find the principal value

$$\text{pr. v.} \int_{-\infty}^{\infty} \frac{dx}{(x^2 - 3x + 2)(x^2 + 1)}.$$

Solution. Since

$$x^2 - 3x + 2 = (x - 1)(x - 2),$$

the integrand $f(x)$, considered for complex z, has simple poles at

$$z = 1, \qquad \operatorname*{Res}_{z=1} f(z) = \left[\frac{1}{(z - 2)(z^2 + 1)}\right]_{z=1} = -\frac{1}{2},$$

$$z = 2, \qquad \operatorname*{Res}_{z=2} f(z) = \left[\frac{1}{(z - 1)(z^2 + 1)}\right]_{z=2} = \frac{1}{5},$$

$$z = i, \qquad \operatorname*{Res}_{z=i} f(z) = \left[\frac{1}{(z^2 - 3z + 2)(z + i)}\right]_{z=i} = \frac{1}{6 + 2i} = \frac{3 - i}{20},$$

and at $z = -i$ in the lower half-plane, which is of no interest here. From (14) we get the answer

$$\text{pr. v.} \int_{-\infty}^{\infty} \frac{dx}{(x^2 - 3x + 2)(x^2 + 1)} = 2\pi i \left(\frac{3 - i}{20}\right) + \pi i \left(-\frac{1}{2} + \frac{1}{5}\right) = \frac{\pi}{10}.$$ ■

More integrals of the kind considered in this section are included in the problem set. Try also your CAS, which may sometimes give you false results on complex integrals.

PROBLEM SET 16.4

1–9 INTEGRALS INVOLVING COSINE AND SINE

Evaluate the following integrals and show the details of your work.

1. $\int_0^{\pi} \frac{2\, d\theta}{k - \cos\theta}$

2. $\int_0^{\pi} \frac{d\theta}{\pi + 3\cos\theta}$

3. $\int_0^{2\pi} \frac{1 + \sin\theta}{3 + \cos\theta}\, d\theta$

4. $\int_0^{2\pi} \frac{1 + 4\cos\theta}{17 - 8\cos\theta}\, d\theta$

5. $\int_0^{2\pi} \frac{\cos^2\theta}{5 - 4\cos\theta}\, d\theta$

6. $\int_0^{2\pi} \frac{\sin^2\theta}{5 - 4\cos\theta}\, d\theta$

7. $\int_0^{2\pi} \frac{a}{a - \sin\theta}\, d\theta$

8. $\int_0^{2\pi} \frac{1}{8 - 2\sin\theta}\, d\theta$

9. $\int_0^{2\pi} \frac{\cos\theta}{13 - 12\cos 2\theta}\, d\theta$

10–22 IMPROPER INTEGRALS: INFINITE INTERVAL OF INTEGRATION

Evaluate the following integrals and show details of your work.

10. $\int_{-\infty}^{\infty} \frac{dx}{(1 + x^2)^3}$

11. $\int_{-\infty}^{\infty} \frac{dx}{(1 + x^2)^2}$

12. $\int_{-\infty}^{\infty} \frac{dx}{(x^2 - 2x + 5)^2}$

13. $\int_{-\infty}^{\infty} \frac{x}{(x^2 + 1)(x^2 + 4)}\, dx$

14. $\int_{-\infty}^{\infty} \frac{x^2 + 1}{x^4 + 1}\, dx$

15. $\int_{-\infty}^{\infty} \frac{x^2}{x^6 + 1}\, dx$

16. $\int_{-\infty}^{\infty} \frac{\cos 2x}{(x^2 + 1)^2}\, dx$

17. $\int_{-\infty}^{\infty} \frac{\sin 3x}{x^4 + 1}\, dx$

18. $\int_{-\infty}^{\infty} \frac{\cos 4x}{x^4 + 5x^2 + 4}\, dx$

19. $\int_{-\infty}^{\infty} \frac{dx}{x^4 - 1}$

20. $\int_{-\infty}^{\infty} \frac{x}{8 - x^3}\, dx$

21. $\int_{-\infty}^{\infty} \frac{\sin x}{(x - 1)(x^2 + 4)}\, dx$

22. $\int_{-\infty}^{\infty} \frac{dx}{x^2 - ix}$

23–26 IMPROPER INTEGRALS: POLES ON THE REAL AXIS

Find the Cauchy principal value (showing details):

23. $\int_{-\infty}^{\infty} \frac{dx}{x^4 - 1}$

24. $\int_{-\infty}^{\infty} \frac{dx}{x^4 + 3x^2 - 4}$

25. $\int_{-\infty}^{\infty} \frac{x + 5}{x^3 - x}\, dx$

26. $\int_{-\infty}^{\infty} \frac{x^2}{x^4 - 1}\, dx$

27. CAS EXPERIMENT. Simple Poles on the Real Axis. Experiment with integrals $\int_{-\infty}^{\infty} f(x)\, dx$, $f(x) = [(x - a_1)(x - a_2) \cdots (x - a_k)]^{-1}$, a_j real and all different, $k > 1$. Conjecture that the principal value of these integrals is 0. Try to prove this for a special k, say, $k = 3$. For general k.

28. TEAM PROJECT. Comments on Real Integrals.

(a) Formula (10) follows from (9). Give the details.

(b) Use of auxiliary results. Integrating e^{-z^2} around the boundary C of the rectangle with vertices $-a$, a, $a + ib$, $-a + ib$, letting $a \to \infty$, and using

$$\int_0^{\infty} e^{-x^2} dx = \frac{\sqrt{\pi}}{2},$$

show that

$$\int_0^{\infty} e^{-x^2} \cos 2bx\, dx = \frac{\sqrt{\pi}}{2} e^{-b^2}.$$

(This integral is needed in heat conduction in Sec. 12.7.)

(c) Inspection. Solve Probs. 13 and 17 without calculation.

CHAPTER 16 REVIEW QUESTIONS AND PROBLEMS

1. What is a Laurent series? Its principal part? Its use? Give simple examples.

2. What kind of singularities did we discuss? Give definitions and examples.

3. What is the residue? Its role in integration? Explain methods to obtain it.

4. Can the residue at a singularity be zero? At a simple pole? Give reason.

5. State the residue theorem and the idea of its proof from memory.

6. How did we evaluate real integrals by residue integration? How did we obtain the closed paths needed?

7. What are improper integrals? Their principal value? Why did they occur in this chapter?

8. What do you know about zeros of analytic functions? Give examples.

9. What is the extended complex plane? The Riemann sphere R? Sketch $z = 1 + i$ on R.

10. What is an entire function? Can it be analytic at infinity? Explain the definitions.

11–18 COMPLEX INTEGRALS

Integrate counterclockwise around C. Show the details.

11. $\dfrac{\sin 3z}{z^2}$, $C\colon |z| = \pi$

12. $e^{2/z}$, $C\colon |z - 1 - i| = 2$

13. $\dfrac{5z^3}{z^2 + 4}$, $C\colon |z| = 3$

14. $\dfrac{5z^3}{z^2 + 4}$, $C\colon |z - i| = \pi i/2$

15. $\dfrac{25z^2}{(z - 5)^2}$, $C\colon |z - 5| = 1$

16. $\dfrac{15z + 9}{z^3 - 9z}$, $C\colon |z| = 4$

17. $\dfrac{\cos z}{z^n}$, $n = 0, 1, 2, \cdots$, $C\colon |z| = 1$

18. $\cot 4z$, $C\colon |z| = \frac{3}{4}$

19–25 REAL INTEGRALS

Evaluate by the methods of this chapter. Show details.

19. $\displaystyle\int_0^{2\pi} \frac{d\theta}{13 - 5\sin\theta}$

20. $\displaystyle\int_0^{2\pi} \frac{\sin\theta}{3 + \cos\theta}\, d\theta$

21. $\displaystyle\int_0^{2\pi} \frac{\sin\theta}{34 - 16\sin\theta}\, d\theta$

22. $\displaystyle\int_{-\infty}^{\infty} \frac{dx}{1 + 4x^4}$

23. $\displaystyle\int_{-\infty}^{\infty} \frac{x}{(1 + x^2)^2}\, dx$

24. $\displaystyle\int_{-\infty}^{\infty} \frac{dx}{x^2 - 4ix}$

25. $\displaystyle\int_{-\infty}^{\infty} \frac{\cos x}{x^2 + 1}\, dx$

SUMMARY OF CHAPTER 16
Laurent Series. Residue Integration

A **Laurent series** is a series of the form

$$f(z) = \sum_{n=0}^{\infty} a_n (z - z_0)^n + \sum_{n=1}^{\infty} \frac{b_n}{(z - z_0)^n} \qquad (1) \quad \text{(Sec. 16.1)}$$

or, more briefly written [but this means the same as (1)!]

$$f(z) = \sum_{n=-\infty}^{\infty} a_n (z - z_0)^n, \qquad a_n = \frac{1}{2\pi i} \oint_C \frac{f(z^*)}{(z^* - z_0)^{n+1}}\, dz^* \qquad (1^*)$$

where $n = 0, \pm 1, \pm 2, \cdots$. This series converges in an open annulus (ring) A with center z_0. In A the function $f(z)$ is analytic. At points not in A it may have singularities. The first series in (1) is a power series. In a given annulus, a Laurent series of $f(z)$ is unique, but $f(z)$ may have different Laurent series in different annuli with the same center.

Of particular importance is the Laurent series (1) that converges in a neighborhood of z_0 except at z_0 itself, say, for $0 < |z - z_0| < R$ ($R > 0$, suitable). The series

(or finite sum) of the negative powers in this Laurent series is called the **principal part** of $f(z)$ at z_0. The coefficient b_1 of $1/(z - z_0)$ in this series is called the **residue** of $f(z)$ at z_0 and is given by [see (1) and (1*)]

$$(2) \quad b_1 = \operatorname*{Res}_{z \to z_0} f(z) = \frac{1}{2\pi i} \oint_C f(z^*)\, dz^*. \qquad \text{Thus} \qquad \oint_C f(z^*)\, dz^* = 2\pi i \operatorname*{Res}_{z=z_0} f(z).$$

b_1 can be used for ***integration*** as shown in (2) because it can be found from

$$(3) \qquad \operatorname*{Res}_{z=z_0} f(z) = \frac{1}{(m-1)!} \lim_{z \to z_0} \left(\frac{d^{m-1}}{dz^{m-1}} [(z - z_0)^m f(z)] \right), \qquad \text{(Sec. 16.3)},$$

provided $f(z)$ has at z_0 a **pole of order *m***; by definition this means that principal part has $1/(z - z_0)^m$ as its highest negative power. Thus for a simple pole ($m = 1$),

$$\operatorname*{Res}_{z=z_0} f(z) = \lim_{z \to z_0} (z - z_0) f(z); \qquad \text{also,} \qquad \operatorname*{Res}_{z=z_0} \frac{p(z)}{q(z)} = \frac{p(z_0)}{q'(z_0)}.$$

If the principal part is an infinite series, the singularity of $f(z)$ at z_0 is called an **essential singularity** (Sec. 16.2).

Section 16.2 also discusses the *extended complex plane*, that is, the complex plane with an improper point ∞ ("infinity") attached.

Residue integration may also be used to evaluate certain classes of complicated real integrals (Sec. 16.4).

CHAPTER 17

Conformal Mapping

Conformal mappings are invaluable to the engineer and physicist as an aid in solving problems in potential theory. They are a standard method for solving *boundary value problems* in two-dimensional potential theory and yield rich applications in electrostatics, heat flow, and fluid flow, as we shall see in Chapter 18.

The main feature of conformal mappings is that they are angle-preserving (except at some critical points) and allow a *geometric approach to complex analysis.* More details are as follows. Consider a complex function $w = f(z)$ defined in a domain D of the z–plane; then to each point in D there corresponds a point in the w-plane. In this way we obtain a **mapping** of D onto the range of values of $f(z)$ in the w-plane. In Sec. 17.1 we show that if $f(z)$ is an analytic function, then the mapping given by $w = f(z)$ is a **conformal mapping**, that is, it preserves angles, except at points where the derivative $f'(z)$ is zero. (Such points are called critical points.)

Conformality appeared early in the history of construction of maps of the globe. Such maps can be either "conformal," that is, give directions correctly, or "equiareal," that is, give areas correctly except for a scale factor. However, the maps will always be distorted because they cannot have both properties, as can be proven, see [GenRef8] in App. 1. The designer of accurate maps then has to select which distortion to take into account.

Our study of conformality is similar to the approach used in calculus where we study properties of real functions $y = f(x)$ and graph them. Here we study the properties of conformal mappings (Secs. 17.1–17.4) to get a deeper understanding of the properties of functions, most notably the ones discussed in Chap. 13. Chapter 17 ends with an introduction to ***Riemann surfaces***, an ingenious geometric way of dealing with multivalued complex functions such as $w = \text{sqrt}\,(z)$ and $w = \ln z$.

So far we have covered two main approaches to solving problems in complex analysis. The first one was solving complex integrals by Cauchy's integral formula and was broadly covered by material in Chaps. 13 and 14. The second approach was to use Laurent series and solve complex integrals by residue integration in Chaps. 15 and 16. Now, in Chaps. 17 and 18, we develop a third approach, that is, the *geometric approach* of conformal mapping to solve boundary value problems in complex analysis.

Prerequisite: Chap. 13.
Sections that may be omitted in a shorter course: 17.3 and 17.5.
References and Answers to Problems: App. 1 Part D, App. 2.

17.1 Geometry of Analytic Functions: Conformal Mapping

We shall see that conformal mappings are those mappings that preserve angles, except at critical points, and that these mappings are defined by analytic functions. A critical point occurs wherever the derivative of such a function is zero. To arrive at these results, we have to define terms more precisely.

A complex function

$$w = f(z) = u(x, y) + iv(x, y) \qquad (z = x + iy) \tag{1}$$

of a complex variable z gives a **mapping** of its domain of definition D in the complex z-plane ***into*** the complex w-plane or ***onto*** its range of values in that plane.[1] For any point z_0 in D the point $w_0 = f(z_0)$ is called the **image** of z_0 with respect to f. More generally, for the points of a curve C in D the image points form the **image** of C; similarly for other point sets in D. Also, instead of *the mapping by a function* $w = f(z)$ we shall say more briefly ***the mapping*** $\boldsymbol{w = f(z)}$.

EXAMPLE 1 **Mapping $w = f(x) = z^2$**

Using polar forms $z = re^{i\theta}$ and $w = Re^{i\phi}$, we have $w = z^2 = r^2 e^{2i\theta}$. Comparing moduli and arguments gives $R = r^2$ and $\phi = 2\theta$. Hence circles $r = r_0$ are mapped onto circles $R = r_0^2$ and rays $\theta = \theta_0$ onto rays $\phi = 2\theta_0$. Figure 378 shows this for the region $1 \leqq |z| \leqq \frac{3}{2}$, $\pi/6 \leqq \theta \leqq \pi/3$, which is mapped onto the region $1 \leqq |w| \leqq \frac{9}{4}$, $\pi/3 \leqq \theta \leqq 2\pi/3$.

In Cartesian coordinates we have $z = x + iy$ and

$$u = \text{Re}\,(z^2) = x^2 - y^2, \qquad v = \text{Im}\,(z^2) = 2xy.$$

Hence vertical lines $x = c = \text{const}$ are mapped onto $u = c^2 - y^2$, $v = 2cy$. From this we can eliminate y. We obtain $y^2 = c^2 - u$ and $v^2 = 4c^2y^2$. Together,

$$v^2 = 4c^2(c^2 - u) \qquad \text{(Fig. 379)}.$$

These parabolas open to the left. Similarly, horizontal lines $y = k = \text{const}$ are mapped onto parabolas opening to the right,

$$v^2 = 4k^2(k^2 + u) \qquad \text{(Fig. 379)}. \quad \blacksquare$$

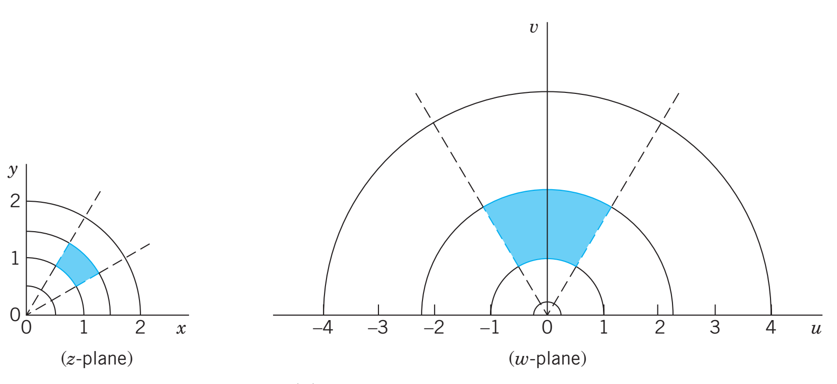


Fig. 378. Mapping $w = z^2$. Lines $|z| = \text{const}$, $\arg z = \text{const}$ and their images in the w-plane

[1]The general terminology is as follows. A mapping of a set A into a set B is called **surjective** or a mapping of A **onto** B if every element of B is the image of at least one element of A. It is called **injective** or **one-to-one** if different elements of A have different images in B. Finally, it is called **bijective** if it is both surjective and injective.

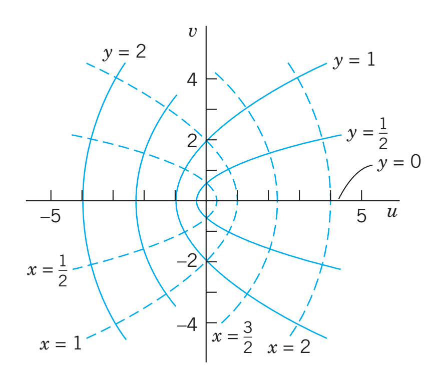


Fig. 379. Images of x = const, y = const under $w = z^2$

Conformal Mapping

A mapping $w = f(z)$ is called **conformal** if it preserves angles between oriented curves in magnitude as well as in sense. Figure 380 shows what this means. The **angle** α $(0 \leqq \alpha \leqq \pi)$ between two intersecting curves C_1 and C_2 is defined to be the angle between their oriented tangents at the intersection point z_0. And *conformality* means that the images C_1^* and C_2^* of C_1 and C_2 make the same angle as the curves themselves in both magnitude and direction.

THEOREM 1

Conformality of Mapping by Analytic Functions

The mapping $w = f(z)$ *by an analytic function* f *is conformal, except at* **critical points**, *that is, points at which the derivative* f' *is zero.*

PROOF $w = z^2$ has a critical point at $z = 0$, where $f'(z) = 2z = 0$ and the angles are doubled (see Fig. 378), so that conformality fails.

The idea of proof is to consider a curve

$$C: \; z(t) = x(t) + iy(t) \tag{2}$$

in the domain of $f(z)$ and to show that $w = f(z)$ rotates all tangents at a point z_0 (where $f'(z_0) \neq 0$) through the same angle. Now $\dot{z}(t) = dz/dt = \dot{x}(t) + i\dot{y}(t)$ is tangent to C in (2) because this is the limit of $(z_1 - z_0)/\Delta t$ (which has the direction of the secant $z_1 - z_0$

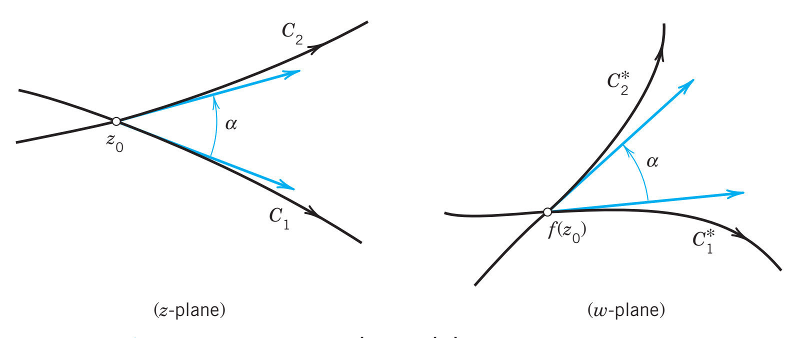


Fig. 380. Curves C_1 and C_2 and their respective images C_1^* and C_2^* under a conformal mapping $w = f(z)$

in Fig. 381) as z_1 approaches z_0 along C. The image C^* of C is $w = f(z(t))$. By the chain rule, $\dot{w} = f'(z(t))\dot{z}(t)$. Hence the tangent direction of C^* is given by the argument (use (9) in Sec. 13.2)

$$\arg \dot{w} = \arg f' + \arg \dot{z} \tag{3}$$

where $\arg \dot{z}$ gives the tangent direction of C. This shows that the mapping rotates *all* directions at a point z_0 in the domain of analyticity of f through the same angle $\arg f'(z_0)$, which exists as long as $f'(z_0) \neq 0$. But this means conformality, as Fig. 381 illustrates for an angle α between two curves, whose images C_1^* and C_2^* make the same angle (because of the rotation). ■

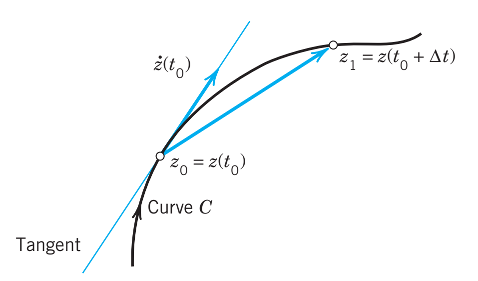


Fig. 381. Secant and tangent of the curve C

In the remainder of this section and in the next ones we shall consider various conformal mappings that are of practical interest, for instance, in modeling potential problems.

EXAMPLE 2 Conformality of $w = z^n$

The mapping $w = z^n$, $n = 2, 3, \cdots$, is conformal, except at $z = 0$, where $w' = nz^{n-1} = 0$. For $n = 2$ this is shown in Fig. 378; we see that at 0 the angles are doubled. For general n the angles at 0 are multiplied by a factor n under the mapping. Hence the sector $0 \leqq \theta \leqq \pi/n$ is mapped by z^n onto the upper half-plane $v \geqq 0$ (Fig. 382). ■

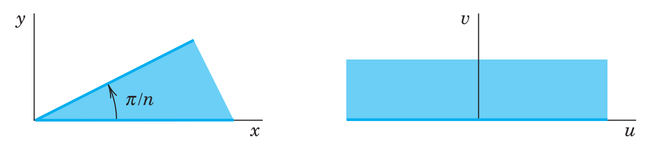


Fig. 382. Mapping by $w = z^n$

EXAMPLE 3 Mapping $w = z + 1/z$. Joukowski Airfoil

In terms of polar coordinates this mapping is

$$w = u + iv = r(\cos\theta + i\sin\theta) + \frac{1}{r}(\cos\theta - i\sin\theta).$$

By separating the real and imaginary parts we thus obtain

$$u = a\cos\theta, \qquad v = b\sin\theta \qquad \text{where} \qquad a = r + \frac{1}{r}, \qquad b = r - \frac{1}{r}.$$

Hence circles $|z| = r = \text{const} \neq 1$ are mapped onto ellipses $x^2/a^2 + y^2/b^2 = 1$. The circle $r = 1$ is mapped onto the segment $-2 \leqq u \leqq 2$ of the u-axis. See Fig. 383.

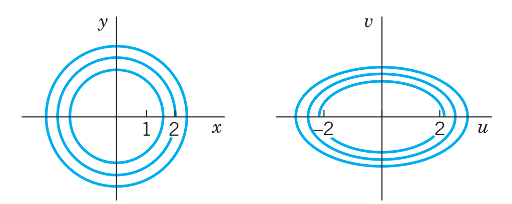


Fig. 383. Example 3

Now the derivative of w is

$$w' = 1 - \frac{1}{z^2} = \frac{(z+1)(z-1)}{z^2}$$

which is 0 at $z = \pm 1$. These are the points at which the mapping is not conformal. The two circles in Fig. 384 pass through $z = -1$. The larger is mapped onto a *Joukowski airfoil*. The dashed circle passes through both -1 and 1 and is mapped onto a curved segment.

Another interesting application of $w = z + 1/z$ (the flow around a cylinder) will be considered in Sec. 18.4. ■

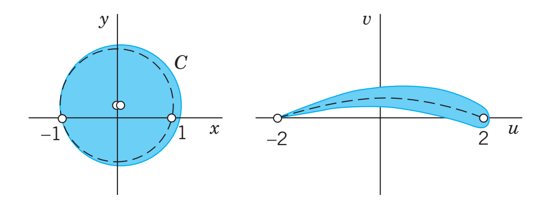


Fig. 384. Joukowski airfoil

EXAMPLE 4 **Conformality of $w = e^z$**

From (10) in Sec. 13.5 we have $|e^z| = e^x$ and $\text{Arg } z = y$. Hence e^z maps a vertical straight line $x = x_0 = \text{const}$ onto the circle $|w| = e^{x_0}$ and a horizontal straight line $y = y_0 = \text{const}$ onto the ray $\arg w = y_0$. The rectangle in Fig. 385 is mapped onto a region bounded by circles and rays as shown.

The fundamental region $-\pi < \text{Arg } z \leqq \pi$ of e^z in the z-plane is mapped bijectively and conformally onto the entire w-plane without the origin $w = 0$ (because $e^z = 0$ for no z). Figure 386 shows that the upper half $0 < y \leqq \pi$ of the fundamental region is mapped onto the upper half-plane $0 < \arg w \leqq \pi$, the left half being mapped inside the unit disk $|w| \leqq 1$ and the right half outside (why?). ■

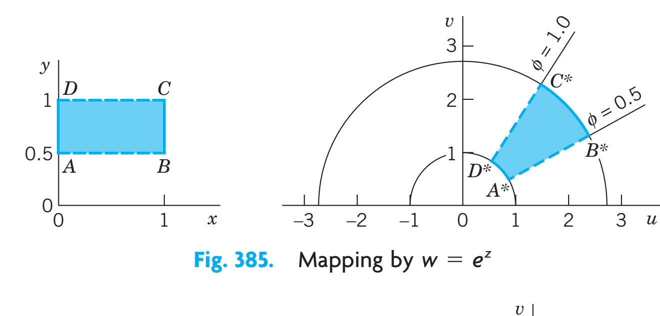


Fig. 385. Mapping by $w = e^z$

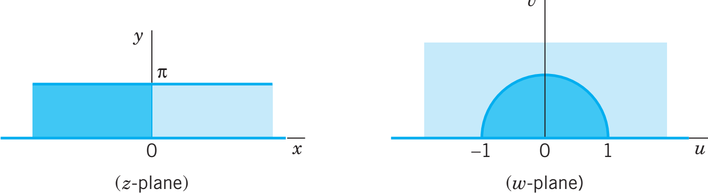


Fig. 386. Mapping by $w = e^z$

EXAMPLE 5 **Principle of Inverse Mapping. Mapping $w = \operatorname{Ln} z$**

Principle. *The mapping by the inverse $z = f^{-1}(w)$ of $w = f(z)$ is obtained by interchanging the roles of the z-plane and the w-plane in the mapping by $w = f(z)$.*

Now the principal value $w = f(z) = \operatorname{Ln} z$ of the natural logarithm has the inverse $z = f^{-1}(w) = e^w$. From Example 4 (with the notations z and w interchanged!) we know that $f^{-1}(w) = e^w$ maps the fundamental region of the exponential function onto the z-plane without $z = 0$ (because $e^w \neq 0$ for every w). Hence $w = f(z) = \operatorname{Ln} z$ maps the z-plane without the origin and cut along the negative real axis (where $\theta = \operatorname{Im} \operatorname{Ln} z$ jumps by 2π) conformally onto the horizontal strip $-\pi < v \leqq \pi$ of the w-plane, where $w = u + iv$.

Since the mapping $w = \operatorname{Ln} z + 2\pi i$ differs from $w = \operatorname{Ln} z$ by the translation $2\pi i$ (vertically upward), this function maps the z-plane (cut as before and 0 omitted) onto the strip $\pi < v \leqq 3\pi$. Similarly for each of the infinitely many mappings $w = \ln z = \operatorname{Ln} z \pm 2n\pi i$ ($n = 0, 1, 2, \cdots$). The corresponding horizontal strips of width 2π (images of the z-plane under these mappings) together cover the whole w-plane without overlapping. ■

Magnification Ratio. By the definition of the derivative we have

$$\lim_{z \to z_0} \left| \frac{f(z) - f(z_0)}{z - z_0} \right| = |f'(z_0)|. \tag{4}$$

Therefore, the mapping $w = f(z)$ magnifies (or shortens) the lengths of short lines by approximately the factor $|f'(z_0)|$. The image of a small figure ***conforms*** to the original figure in the sense that it has approximately the same shape. However, since $f'(z)$ varies from point to point, a *large* figure may have an image whose shape is quite different from that of the original figure.

More on the Condition $f'(z) \neq 0$. From (4) in Sec. 13.4 and the Cauchy–Riemann equations we obtain

$$|f'(z)|^2 = \left| \frac{\partial u}{\partial x} + i \frac{\partial v}{\partial x} \right|^2 = \left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial v}{\partial x}\right)^2 = \frac{\partial u}{\partial x}\frac{\partial v}{\partial y} - \frac{\partial u}{\partial y}\frac{\partial v}{\partial x} \tag{5'}$$

that is,

$$|f'(z)|^2 = \begin{vmatrix} \dfrac{\partial u}{\partial x} & \dfrac{\partial u}{\partial y} \\ \dfrac{\partial v}{\partial x} & \dfrac{\partial v}{\partial y} \end{vmatrix} = \frac{\partial(u, v)}{\partial(x, y)}. \tag{5}$$

This determinant is the so-called **Jacobian** (Sec. 10.3) of the transformation $w = f(z)$ written in real form $u = u(x, y)$, $v = v(x, y)$. Hence $f'(z_0) \neq 0$ implies that the Jacobian is not 0 at z_0. This condition is sufficient that the mapping $w = f(z)$ in a sufficiently small neighborhood of z_0 is one-to-one or injective (different points have different images). See Ref. [GenRef4] in App. 1.

PROBLEM SET 17.1

1. **On Fig. 378.** One "rectangle" and its image are colored. Identify the images for the other "rectangles."
2. **On Example 1.** Verify all calculations.
3. **Mapping $w = z^3$.** Draw an analog of Fig. 378 for $w = z^3$.
4. **Conformality.** Why do the images of the straight lines $x = \text{const}$ and $y = \text{const}$ under a mapping by an analytic function intersect at right angles? Same question for the curves $|z| = \text{const}$ and $\arg z = \text{const}$. Are there exceptional points?

5. **Experiment on $w = \bar{z}$.** Find out whether $w = \bar{z}$ preserves angles in size as well as in sense. Try to prove your result.

6–9 MAPPING OF CURVES

Find and sketch or graph the images of the given curves under the given mapping.

6. $x = 1, 2, 3, 4, \quad y = 1, 2, 3, 4, \quad w = z^2$
7. **Rotation.** Curves as in Prob. 6, $w = iz$
8. **Reflection in the unit circle.** $|z| = \frac{1}{3}, \frac{1}{2}, 1, 2, 3,$ $\text{Arg } z = 0, \pm\pi/4, \pm\pi/2, \pm 3\pi/2$
9. **Translation.** Curves as in Prob. 6, $w = z + 2 + i$
10. **CAS EXPERIMENT. Orthogonal Nets.** Graph the orthogonal net of the two families of level curves $\text{Re} f(z) = \text{const}$ and $\text{Im} f(z) = \text{const}$, where **(a)** $f(z) = z^4$, **(b)** $f(z) = 1/z$, **(c)** $f(z) = 1/z^2$, **(d)** $f(z) = (z + i)/(1 + iz)$. Why do these curves generally intersect at right angles? In your work, experiment to get the best possible graphs. Also do the same for other functions of your own choice. Observe and record shortcomings of your CAS and means to overcome such deficiencies.

11–20 MAPPING OF REGIONS

Sketch or graph the given region and its image under the given mapping.

11. $|z| \leqq \frac{1}{2}, \quad -\pi/8 < \text{Arg } z < \pi/8, \quad w = z^2$
12. $1 < |z| < 3, \quad 0 < \text{Arg } z < \pi/2, \quad w = z^3$
13. $2 \leqq \text{Im } z \leqq 5, \quad w = iz$
14. $x \geqq 1, \quad w = 1/z$
15. $|z - \frac{1}{2}| \leqq \frac{1}{2}, \quad w = 1/z$
16. $|z| < \frac{1}{2}, \quad \text{Im } z > 0, \quad w = 1/z$
17. $-\text{Ln } 2 \leqq x \leqq \text{Ln } 4, \quad w = e^z$
18. $-1 \leqq x \leqq 2, \quad -\pi < y < \pi, \quad w = e^z$
19. $1 < |z| < 4, \quad \pi/4 < \theta \leqq 3\pi/4, \quad w = \text{Ln } z$
20. $\frac{1}{2} \leqq |z| \leqq 1, \quad 0 \leqq \theta < \pi/2, \quad w = \text{Ln } z$

21–26 FAILURE OF CONFORMALITY

Find all points at which the mapping is not conformal. Give reason.

21. A cubic polynomial
22. $z^2 + 1/z^2$
23. $\dfrac{z + \frac{1}{2}}{4z^2 + 2}$
24. $\exp(z^5 - 80z)$
25. $\cosh z$
26. $\sin \pi z$
27. **Magnification of Angles.** Let $f(z)$ be analytic at z_0. Suppose that $f'(z_0) = 0, \cdots, f^{(k-1)}(z_0) = 0$. Then the mapping $w = f(z)$ magnifies angles with vertex at z_0 by a factor k. Illustrate this with examples for $k = 2, 3, 4$.
28. Prove the statement in Prob. 27 for general $k = 1, 2, \cdots$. *Hint.* Use the Taylor series.

29–35 MAGNIFICATION RATIO, JACOBIAN

Find the magnification ratio M. Describe what it tells you about the mapping. Where is $M = 1$? Find the Jacobian J.

29. $w = \frac{1}{2} z^2$
30. $w = z^3$
31. $w = 1/z$
32. $w = 1/z^2$
33. $w = e^z$
34. $w = \dfrac{z + 1}{2z - 2}$
35. $w = \text{Ln } z$

17.2 Linear Fractional Transformations (Möbius Transformations)

Conformal mappings can help in modeling and solving boundary value problems by first mapping regions conformally onto another. We shall explain this for standard regions (disks, half-planes, strips) in the next section. For this it is useful to know properties of special basic mappings. Accordingly, let us begin with the following very important class.

The next two sections discuss linear fractional transformations. The reason for our thorough study is that such transformations are useful in modeling and solving boundary value problems, as we shall see in Chapter 18. The task is to get a good grasp of which

conformal mappings map certain regions conformally onto each other, such as, say mapping a disk onto a half-plane (Sec. 17.3) and so forth. Indeed, the first step in the modeling process of solving boundary value problems is to identify the correct conformal mapping that is related to the "geometry" of the boundary value problem.

The following class of conformal mappings is very important. **Linear fractional transformations** (or **Möbius transformations**) are mappings

$$w = \frac{az + b}{cz + d} \qquad (ad - bc \neq 0) \tag{1}$$

where a, b, c, d are complex or real numbers. Differentiation gives

$$w' = \frac{a(cz + d) - c(az + b)}{(cz + d)^2} = \frac{ad - bc}{(cz + d)^2}. \tag{2}$$

This motivates our requirement $ad - bc \neq 0$. It implies conformality for all z and excludes the totally uninteresting case $w' \equiv 0$ once and for all. Special cases of (1) are

$$\begin{aligned} &w = z + b && (\textit{Translations}) \\ &w = az \quad \text{with } |a| = 1 && (\textit{Rotations}) \\ &w = az + b && (\textit{Linear transformations}) \\ &w = 1/z && (\textit{Inversion in the unit circle}). \end{aligned} \tag{3}$$

EXAMPLE 1 Properties of the Inversion $w = 1/z$ (Fig. 387)

In polar forms $z = re^{i\theta}$ and $w = Re^{i\phi}$ the inversion $w = 1/z$ is

$$Re^{i\phi} = \frac{1}{re^{i\theta}} = \frac{1}{r}e^{-i\theta} \qquad \text{and gives} \qquad R = \frac{1}{r}, \quad \phi = -\theta.$$

Hence the unit circle $|z| = r = 1$ is mapped onto the unit circle $|w| = R = 1$; $w = e^{i\phi} = e^{-i\theta}$. For a general z the image $w = 1/z$ can be found geometrically by marking $|w| = R = 1/r$ on the segment from 0 to z and then reflecting the mark in the real axis. (Make a sketch.)

Figure 387 shows that $w = 1/z$ maps horizontal and vertical straight lines onto circles or straight lines. Even the following is true.

$w = 1/z$ maps every straight line or circle onto a circle or straight line.

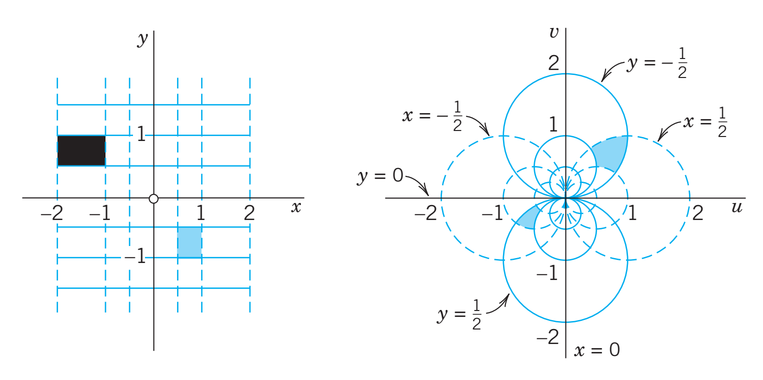


Fig. 387. Mapping (Inversion) $w = 1/z$

Proof. Every straight line or circle in the z-plane can be written

$$A(x^2 + y^2) + Bx + Cy + D = 0 \qquad (A, B\ C, D \text{ real}).$$

$A = 0$ gives a straight line and $A \neq 0$ a circle. In terms of z and $\bar{z}$ this equation becomes

$$Az\bar{z} + B\frac{z + \bar{z}}{2} + C\frac{z - \bar{z}}{2i} + D = 0.$$

Now $w = 1/z$. Substitution of $z = 1/w$ and multiplication by $w\bar{w}$ gives the equation

$$A + B\frac{\bar{w} + w}{2} + C\frac{\bar{w} - w}{2i} + Dw\bar{w} = 0$$

or, in terms of u and v,

$$A + Bu - Cv + D(u^2 + v^2) = 0.$$

This represents a circle (if $D \neq 0$) or a straight line (if $D = 0$) in the w-plane. ∎

The proof in this example suggests the use of z and $\bar{z}$ instead of x and y, a ***general principle*** that is often quite useful in practice.

Surprisingly, *every* linear fractional transformation has the property just proved:

THEOREM 1

Circles and Straight Lines

Every linear fractional transformation (1) *maps the totality of circles and straight lines in the z-plane onto the totality of circles and straight lines in the w-plane.*

PROOF This is trivial for a translation or rotation, fairly obvious for a uniform expansion or contraction, and true for $w = 1/z$, as just proved. Hence it also holds for composites of these special mappings. Now comes the key idea of the proof: represent (1) in terms of these special mappings. When $c = 0$, this is easy. When $c \neq 0$, the representation is

$$w = K\frac{1}{cz + d} + \frac{a}{c} \qquad \text{where} \qquad K = -\frac{ad - bc}{c}.$$

This can be verified by substituting K, taking the common denominator and simplifying; this yields (1). We can now set

$$w_1 = cz, \qquad w_2 = w_1 + d, \qquad w_3 = \frac{1}{w_2}, \qquad w_4 = Kw_3,$$

and see from the previous formula that then $w = w_4 + a/c$. This tells us that (1) is indeed a composite of those special mappings and completes the proof. ∎

Extended Complex Plane

The extended complex plane (the complex plane together with the point ∞ in Sec. 16.2) can now be motivated even more naturally by linear fractional transformations as follows.

To each z for which $cz + d \neq 0$ there corresponds a unique w in (1). Now let $c \neq 0$. Then for $z = -d/c$ we have $cz + d = 0$, so that no w corresponds to this z. This suggests that we let $w = \infty$ be the image of $z = -d/c$.

Also, the **inverse mapping** of (1) is obtained by solving (1) for z; this gives again a linear fractional transformation

$$z = \frac{dw - b}{-cw + a}. \tag{4}$$

When $c \neq 0$, then $cw - a = 0$ for $w = a/c$, and we let a/c be the image of $z = \infty$. With these settings, the linear fractional transformation (1) is now a one-to-one mapping of the extended z-plane onto the extended w-plane. We also say that every linear fractional transformation maps "the extended complex plane in a one-to-one manner onto itself."

Our discussion suggests the following.

General Remark. If $z = \infty$, then the right side of (1) becomes the meaningless expression $(a \cdot \infty + b)/(c \cdot \infty + d)$. We assign to it the value $w = a/c$ if $c \neq 0$ and $w = \infty$ if $c = 0$.

Fixed Points

Fixed points of a mapping $w = f(z)$ are points that are mapped onto themselves, are "kept fixed" under the mapping. Thus they are obtained from

$$w = f(z) = z.$$

The **identity mapping** $w = z$ has every point as a fixed point. The mapping $w = \bar{z}$ has infinitely many fixed points, $w = 1/z$ has two, a rotation has one, and a translation none in the finite plane. (Find them in each case.) For (1), the fixed-point condition $w = z$ is

$$z = \frac{az + b}{cz + d}, \qquad \text{thus} \qquad cz^2 - (a - d)z - b = 0. \tag{5}$$

For $c \neq 0$ this is a quadratic equation in z whose coefficients all vanish if and only if the mapping is the identity mapping $w = z$ (in this case, $a = d \neq 0$, $b = c = 0$). Hence we have

THEOREM 2

Fixed Points

A linear fractional transformation, not the identity, has at most two fixed points. If a linear fractional transformation is known to have three or more fixed points, it must be the identity mapping $w = z$.

To make our present general discussion of linear fractional transformations even more useful from a practical point of view, we extend it by further facts and typical examples, in the problem set as well as in the next section.

PROBLEM SET 17.2

1. Verify the calculations in the proof of Theorem 1, including those for the case $c = 0$.

2. Composition of LFTs. Show that substituting a linear fractional transformation (LFT) into an LFT gives an LFT.

3. Matrices. If you are familiar with 2×2 matrices, prove that the coefficient matrices of (1) and (4) are inverses of each other, provided that $ad - bc = 1$, and that the composition of LFTs corresponds to the multiplication of the coefficient matrices.

4. Fig. 387. Find the image of $x = k =$ const under $w = 1/z$. *Hint.* Use formulas similar to those in Example 1.

5. Inverse. Derive (4) from (1) and conversely.

6. Fixed points. Find the fixed points mentioned in the text before formula (5).

7–10 INVERSE

Find the inverse $z = z(w)$. Check by solving $z(w)$ for w.

7. $w = \dfrac{i}{2z - 1}$

8. $w = \dfrac{z - i}{z + i}$

9. $w = \dfrac{z - i}{3iz + 4}$

10. $w = \dfrac{z - \frac{1}{2}i}{-\frac{1}{2}iz - 1}$

11–16 FIXED POINTS

Find the fixed points.

11. $w = (a + ib)z^2$

12. $w = z - 3i$

13. $w = 16z^5$

14. $w = az + b$

15. $w = \dfrac{iz + 4}{2z - 5i}$

16. $w = \dfrac{aiz - 1}{z + ai}, \quad a \neq 1$

17–20 FIXED POINTS

Find all LFTs with fixed point(s).

17. $z = 0$

18. $z = \pm 1$

19. $z = \pm i$

20. Without any fixed points

17.3 Special Linear Fractional Transformations

We continue our study of linear fractional transformations. We shall identify linear fractional transformations

$$w = \frac{az + b}{cz + d} \qquad (ad - bc \neq 0) \tag{1}$$

that map certain standard domains onto others. Theorem 1 (below) will give us a tool for constructing desired linear fractional transformations.

A mapping (1) is determined by a, b, c, d, actually by the ratios of three of these constants to the fourth because we can drop or introduce a common factor. This makes it plausible that three conditions determine a unique mapping (1):

THEOREM 1

Three Points and Their Images Given

Three given distinct points z_1, z_2, z_3 can always be mapped onto three prescribed distinct points w_1, w_2, w_3 by one, and only one, linear fractional transformation $w = f(z)$. This mapping is given implicitly by the equation

$$\frac{w - w_1}{w - w_3} \cdot \frac{w_2 - w_3}{w_2 - w_1} = \frac{z - z_1}{z - z_3} \cdot \frac{z_2 - z_3}{z_2 - z_1}. \tag{2}$$

(If one of these points is the point ∞, the quotient of the two differences containing this point must be replaced by 1.)

PROOF Equation (2) is of the form $F(w) = G(z)$ with linear fractional F and G. Hence $w = F^{-1}(G(z)) = f(z)$, where F^{-1} is the inverse of F and is linear fractional (see (4) in

Sec. 17.2) and so is the composite $F^{-1}(G(z))$ (by Prob. 2 in Sec. 17.2), that is, $w = f(z)$ is linear fractional. Now if in (2) we set $w = w_1, w_2, w_3$ on the left and $z = z_1, z_2, z_3$ on the right, we see that

$$F(w_1) = 0, \qquad F(w_2) = 1, \qquad F(w_3) = \infty$$

$$G(z_1) = 0, \qquad G(z_2) = 1, \qquad G(z_3) = \infty.$$

From the first column, $F(w_1) = G(z_1)$, thus $w_1 = F^{-1}(G(z_1)) = f(z_1)$. Similarly, $w_2 = f(z_2)$, $w_3 = f(z_3)$. This proves the existence of the desired linear fractional transformation.

To prove uniqueness, let $w = g(z)$ be a linear fractional transformation, which also maps z_j onto w_j, $j = 1, 2, 3$. Thus $w_j = g(z_j)$. Hence $g^{-1}(w_j) = z_j$, where $w_j = f(z_j)$. Together, $g^{-1}(f(z_j)) = z_j$, a mapping with the three fixed points z_1, z_2, z_3. By Theorem 2 in Sec. 17.2, this is the identity mapping, $g^{-1}(f(z)) = z$ for all z. Thus $f(z) = g(z)$ for all z, the uniqueness.

The last statement of Theorem 1 follows from the General Remark in Sec. 17.2. ■

Mapping of Standard Domains by Theorem 1

Using Theorem 1, we can now find linear fractional transformations of some practically useful domains (here called "standard domains") according to the following principle.

Principle. Prescribe three boundary points z_1, z_2, z_3 of the domain D in the z-plane. Choose their images w_1, w_2, w_3 on the boundary of the image D^* of D in the w-plane. Obtain the mapping from (2). Make sure that D is mapped onto D^*, not onto its complement. In the latter case, interchange two w-points. (Why does this help?)

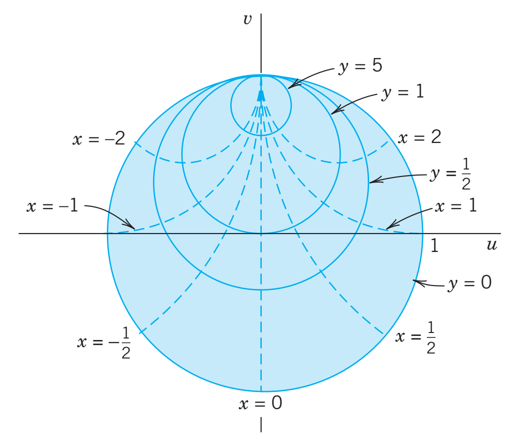


Fig. 388. Linear fractional transformation in Example 1

EXAMPLE 1 Mapping of a Half-Plane onto a Disk (Fig. 388)

Find the linear fractional transformation (1) that maps $z_1 = -1, z_2 = 0, z_3 = 1$ onto $w_1 = -1, w_2 = -i$, $w_3 = 1$, respectively.

Solution. From (2) we obtain

$$\frac{w-(-1)}{w-1} \cdot \frac{-i-1}{-i-(-1)} = \frac{z-(-1)}{z-1} \cdot \frac{0-1}{0-(-1)},$$

thus

$$w = \frac{z - i}{-iz + 1}.$$

Let us show that we can determine the specific properties of such a mapping without much calculation. For $z = x$ we have $w = (x - i)/(-ix + 1)$, thus $|w| = 1$, so that the x-axis maps onto the unit circle. Since $z = i$ gives $w = 0$, the upper half-plane maps onto the interior of that circle and the lower half-plane onto the exterior. $z = 0, i, \infty$ go onto $w = -i, 0, i$, so that the positive imaginary axis maps onto the segment S: $u = 0, -1 \leqq v \leqq 1$. The vertical lines $x = \text{const}$ map onto circles (by Theorem 1, Sec. 17.2) through $w = i$ (the image of $z = \infty$) and perpendicular to $|w| = 1$ (by conformality; see Fig. 388). Similarly, the horizontal lines $y = \text{const}$ map onto circles through $w = i$ and perpendicular to S (by conformality). Figure 388 gives these circles for $y \geqq 0$, and for $y < 0$ they lie outside the unit disk shown. ■

EXAMPLE 2 Occurrence of ∞

Determine the linear fractional transformation that maps $z_1 = 0, z_2 = 1, z_3 = \infty$ onto $w_1 = -1, w_2 = -i, w_3 = 1$, respectively.

Solution. From (2) we obtain the desired mapping

$$w = \frac{z - i}{z + i}.$$

This is sometimes called the *Cayley transformation.*[2] In this case, (2) gave at first the quotient $(1 - \infty)/(z - \infty)$, which we had to replace by 1. ■

EXAMPLE 3 Mapping of a Disk onto a Half-Plane

Find the linear fractional transformation that maps $z_1 = -1, z_2 = i, z_3 = 1$ onto $w_1 = 0, w_2 = i, w_3 = \infty$, respectively, such that the unit disk is mapped onto the right half-plane. (Sketch disk and half-plane.)

Solution. From (2) we obtain, after replacing $(i - \infty)/(w - \infty)$ by 1,

$$w = -\frac{z + 1}{z - 1}.$$ ■

Mapping half-planes onto half-planes is another task of practical interest. For instance, we may wish to map the upper half-plane $y \geqq 0$ onto the upper half-plane $v \geqq 0$. Then the x-axis must be mapped onto the u-axis.

EXAMPLE 4 Mapping of a Half-Plane onto a Half-Plane

Find the linear fractional transformation that maps $z_1 = -2, z_2 = 0, z_3 = 2$ onto $w_1 = \infty, w_2 = \frac{1}{4}, w_3 = \frac{3}{8}$, respectively.

Solution. You may verify that (2) gives the mapping function

$$w = \frac{z + 1}{2z + 4}$$

What is the image of the x-axis? Of the y-axis? ■

Mappings of disks onto disks is a third class of practical problems. We may readily verify that the unit disk in the z-plane is mapped onto the unit disk in the w-plane by the following function, which maps z_0 onto the center $w = 0$.

[2]ARTHUR CAYLEY (1821–1895), English mathematician and professor at Cambridge, is known for his important work in algebra, matrix theory, and differential equations.

$$w = \frac{z - z_0}{cz - 1}, \qquad c = \bar{z}_0, \qquad |z_0| < 1. \tag{3}$$

To see this, take $|z| = 1$, obtaining, with $c = \bar{z}_0$ as in (3),

$$\begin{aligned} |z - z_0| &= |\bar{z} - c| \\ &= |z|\,|\bar{z} - c| \\ &= |z\bar{z} - cz| = |1 - cz| = |cz - 1|. \end{aligned}$$

Hence

$$|w| = |z - z_0|/|cz - 1| = 1$$

from (3), so that $|z| = 1$ maps onto $|w| = 1$, as claimed, with z_0 going onto 0, as the numerator in (3) shows.

Formula (3) is illustrated by the following example. Another interesting case will be given in Prob. 17 of Sec. 18.2.

EXAMPLE 5 Mapping of the Unit Disk onto the Unit Disk

Taking $z_0 = \frac{1}{2}$ in (3), we obtain (verify!)

$$w = \frac{2z - 1}{z - 2} \qquad \text{(Fig. 389).}$$ ■

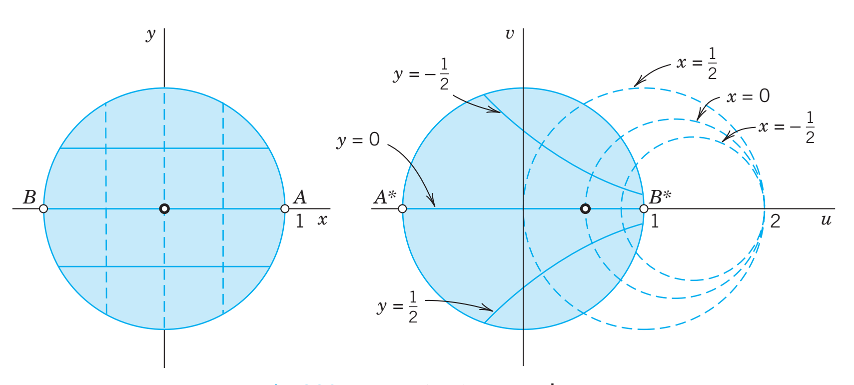


Fig. 389. Mapping in Example 5

EXAMPLE 6 Mapping of an Angular Region onto the Unit Disk

Certain mapping problems can be solved by combining linear fractional transformations with others. For instance, to map the angular region D: $-\pi/6 \leqq \arg z \leqq \pi/6$ (Fig. 390) onto the unit disk $|w| \leqq 1$, we may map D by $Z = z^3$ onto the right Z-half-plane and then the latter onto the disk $|w| \leqq 1$ by

$$w = i\,\frac{Z - 1}{Z + 1}, \qquad \text{combined} \qquad w = i\,\frac{z^3 - 1}{z^3 + 1}.$$ ■

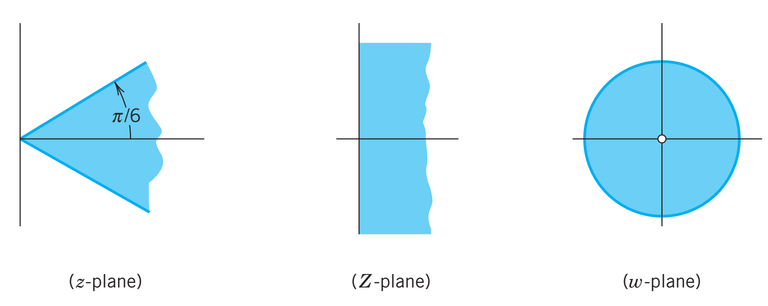


Fig. 390. Mapping in Example 6

This is the end of our discussion of linear fractional transformations. In the next section we turn to conformal mappings by other analytic functions (sine, cosine, etc.).

PROBLEM SET 17.3

1. **CAS EXPERIMENT. Linear Fractional Transformations (LFTs). (a)** Graph typical regions (squares, disks, etc.) and their images under the LFTs in Examples 1–5 of the text.

   (b) Make an experimental study of the continuous dependence of LFTs on their coefficients. For instance, change the LFT in Example 4 continuously and graph the changing image of a fixed region (applying animation if available).
2. **Inverse.** Find the inverse of the mapping in Example 1. Show that under that inverse the lines $x = \text{const}$ are the images of circles in the w-plane with centers on the line $v = 1$.
3. **Inverse.** If $w = f(z)$ is any transformation that has an inverse, prove the (trivial!) fact that f and its inverse have the same fixed points.
4. Obtain the mapping in Example 1 of this section from Prob. 18 in Problem Set 17.2.
5. Derive the mapping in Example 2 from (2).
6. Derive the mapping in Example 4 from (2). Find its inverse and the fixed points.
7. Verify the formula for disks.

8–16 LFTs FROM THREE POINTS AND IMAGES

Find the LFT that maps the given three points onto the three given points in the respective order.

8. $0, 1, 2$ onto $1, \frac{1}{2}, \frac{1}{3}$
9. $1, i, -1$ onto $i, -1, -i$
10. $0, -i, i$ onto $-1, 0, \infty$
11. $-1, 0, 1$ onto $-i, -1, i$
12. $0, 2i, -2i$ onto $-1, 0, \infty$
13. $0, 1, \infty$ onto $\infty, 1, 0$
14. $-1, 0, 1$ onto $1, 1+i, 1+2i$
15. $1, i, 2$ onto $0, -i-1, -\frac{1}{2}$
16. $-\frac{3}{2}, 0, 1$ onto $0, \frac{3}{2}, 1$
17. Find an LFT that maps $|z| \leqq 1$ onto $|w| \leqq 1$ so that $z = i/2$ is mapped onto $w = 0$. Sketch the images of the lines $x = \text{const}$ and $y = \text{const}$.
18. Find all LFTs $w(z)$ that map the x-axis onto the u-axis.
19. Find an analytic function $w = f(z)$ that maps the region $0 \leqq \arg z \leqq \pi/4$ onto the unit disk $|w| \leqq 1$.
20. Find an analytic function that maps the second quadrant of the z-plane onto the interior of the unit circle in the w-plane.

17.4 Conformal Mapping by Other Functions

We shall now cover mappings by trigonometric and hyperbolic analytic functions. So far we have covered the mappings by z^n and e^z (Sec. 17.1) as well as linear fractional transformations (Secs. 17.2 and 17.3).

Sine Function. Figure 391 shows the mapping by

$$w = u + iv = \sin z = \sin x \cosh y + i \cos x \sinh y \qquad \text{(Sec. 13.6)}. \tag{1}$$

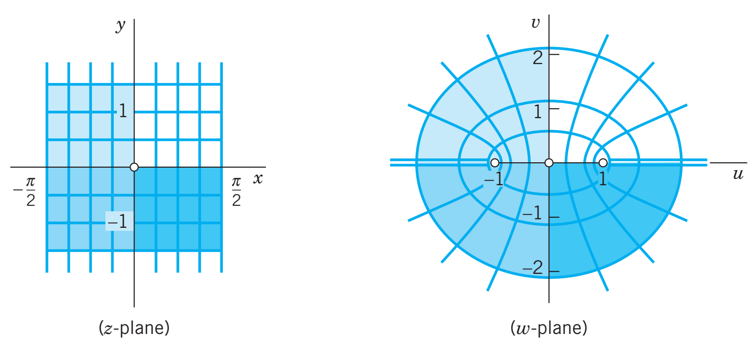


Fig. 391. Mapping $w = u + iv = \sin z$

Hence

$$u = \sin x \cosh y, \qquad v = \cos x \sinh y. \tag{2}$$

Since $\sin z$ is periodic with period 2π, the mapping is certainly not one-to-one if we consider it in the full z-plane. We restrict z to the vertical strip S: $-\frac{1}{2}\pi \leqq x \leqq \frac{1}{2}\pi$ in Fig. 391. Since $f'(z) = \cos z = 0$ at $z = \pm\frac{1}{2}\pi$, the mapping is not conformal at these two critical points. We claim that the rectangular net of straight lines $x = \text{const}$ and $y = \text{const}$ in Fig. 391 is mapped onto a net in the w-plane consisting of hyperbolas (the images of the vertical lines $x = \text{const}$) and ellipses (the images of the horizontal lines $y = \text{const}$) intersecting the hyperbolas at right angles (conformality!). Corresponding calculations are simple. From (2) and the relations $\sin^2 x + \cos^2 x = 1$ and $\cosh^2 y - \sinh^2 y = 1$ we obtain

$$\frac{u^2}{\sin^2 x} - \frac{v^2}{\cos^2 x} = \cosh^2 y - \sinh^2 y = 1 \qquad \text{(Hyperbolas)}$$

$$\frac{u^2}{\cosh^2 y} + \frac{v^2}{\sinh^2 y} = \sin^2 x + \cos^2 x = 1 \qquad \text{(Ellipses).}$$

Exceptions are the vertical lines $x = -\frac{1}{2}\pi x = \frac{1}{2}\pi$, which are "folded" onto $u \leqq -1$ and $u \geqq 1$ $(v = 0)$, respectively.

Figure 392 illustrates this further. The upper and lower sides of the rectangle are mapped onto semi-ellipses and the vertical sides onto $-\cosh 1 \leqq u \leqq -1$ and $1 \leqq u \leqq \cosh 1$ $(v = 0)$, respectively. An application to a potential problem will be given in Prob. 3 of Sec. 18.2.

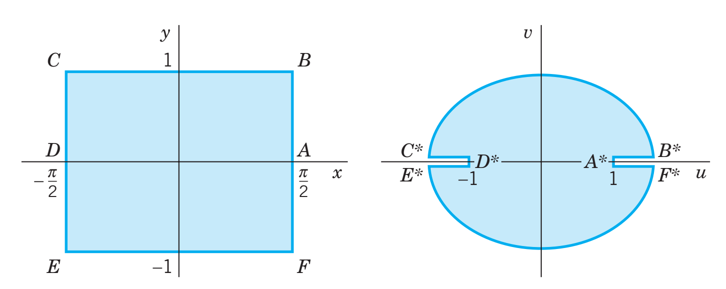


Fig. 392. Mapping by $w = \sin z$

Cosine Function. The mapping $w = \cos z$ could be discussed independently, but since

$$w = \cos z = \sin\ (z + \tfrac{1}{2}\pi), \tag{3}$$

we see at once that this is the same mapping as $\sin z$ preceded by a translation to the right through $\frac{1}{2}\pi$ units.

Hyperbolic Sine. Since

$$w = \sinh z = -i \sin (iz), \tag{4}$$

the mapping is a counterclockwise rotation $Z = iz$ through $\frac{1}{2}\pi$ (i.e., 90°), followed by the sine mapping $Z^* = \sin Z$, followed by a clockwise 90°-rotation $w = -iZ^*$.

Hyperbolic Cosine. This function

$$w = \cosh z = \cos (iz) \tag{5}$$

defines a mapping that is a rotation $Z = iz$ followed by the mapping $w = \cos Z$.

Figure 393 shows the mapping of a semi-infinite strip onto a half-plane by $w = \cosh z$. Since $\cosh 0 = 1$, the point $z = 0$ is mapped onto $w = 1$. For real $z = x \geqq 0$, $\cosh z$ is real and increases with increasing x in a monotone fashion, starting from 1. Hence the positive x-axis is mapped onto the portion $u \geqq 1$ of the u-axis.

For pure imaginary $z = iy$ we have $\cosh iy = \cos y$. Hence the left boundary of the strip is mapped onto the segment $1 \geqq u \geqq -1$ of the u-axis, the point $z = \pi i$ corresponding to

$$w = \cosh i\pi = \cos \pi = -1.$$

On the upper boundary of the strip, $y = \pi$, and since $\sin \pi = 0$ and $\cos \pi = -1$, it follows that this part of the boundary is mapped onto the portion $u \leqq -1$ of the u-axis. Hence the boundary of the strip is mapped onto the u-axis. It is not difficult to see that the interior of the strip is mapped onto the upper half of the w-plane, and the mapping is one-to-one.

This mapping in Fig. 393 has applications in potential theory, as we shall see in Prob. 12 of Sec. 18.3.

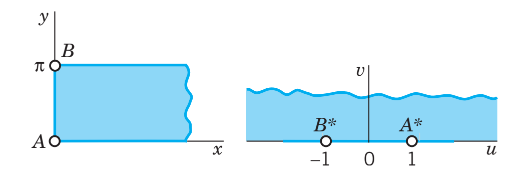


Fig. 393. Mapping by $w = \cosh z$

Tangent Function. Figure 394 shows the mapping of a vertical infinite strip onto the unit circle by $w = \tan z$, accomplished in three steps as suggested by the representation (Sec. 13.6)

$$w = \tan z = \frac{\sin z}{\cos z} = \frac{(e^{iz} - e^{-iz})/i}{e^{iz} + e^{-iz}} = \frac{(e^{2iz} - 1)/i}{e^{2iz} + 1}.$$

Hence if we set $Z = e^{2iz}$ and use $1/i = -i$, we have

$$(6) \qquad w = \tan z = -iW, \qquad W = \frac{Z-1}{Z+1}, \qquad Z = e^{2iz}.$$

We now see that $w = \tan z$ is a linear fractional transformation preceded by an exponential mapping (see Sec. 17.1) and followed by a clockwise rotation through an angle $\frac{1}{2}\pi(90°)$.

The strip is $S: -\frac{1}{4}\pi < x < \frac{1}{4}\pi$, and we show that it is mapped onto the unit disk in the w-plane. Since $Z = e^{2iz} = e^{-2y+2ix}$, we see from (10) in Sec. 13.5 that $|Z| = e^{-2y}$, $\operatorname{Arg} Z = 2x$. Hence the vertical lines $x = -\pi/4, 0, \pi/4$ are mapped onto the rays $\operatorname{Arg} Z = -\pi/2, 0, \pi/2$, respectively. Hence S is mapped onto the right Z-half-plane. Also $|Z| = e^{-2y} < 1$ if $y > 0$ and $|Z| > 1$ if $y < 0$. Hence the upper half of S is mapped inside the unit circle $|Z| = 1$ and the lower half of S outside $|Z| = 1$, as shown in Fig. 394.

Now comes the linear fractional transformation in (6), which we denote by $g(Z)$:

$$(7) \qquad W = g(Z) = \frac{Z-1}{Z+1}.$$

For real Z this is real. Hence the real Z-axis is mapped onto the real W-axis. Furthermore, the imaginary Z-axis is mapped onto the unit circle $|W| = 1$ because for pure imaginary $Z = iY$ we get from (7)

$$|W| = |g(iY)| = \left|\frac{iY-1}{iY+1}\right| = 1.$$

The right Z-half-plane is mapped inside this unit circle $|W| = 1$, not outside, because $Z = 1$ has its image $g(1) = 0$ inside that circle. Finally, the unit circle $|Z| = 1$ is mapped onto the imaginary W-axis, because this circle is $Z = e^{i\phi}$, so that (7) gives a pure imaginary expression, namely,

$$g(e^{i\phi}) = \frac{e^{i\phi}-1}{e^{i\phi}+1} = \frac{e^{i\phi/2}-e^{-i\phi/2}}{e^{i\phi/2}+e^{-i\phi/2}} = \frac{i\sin(\phi/2)}{\cos(\phi/2)}.$$

From the W-plane we get to the w-plane simply by a clockwise rotation through $\pi/2$; see (6).

Together we have shown that $w = \tan z$ maps $S: -\pi/4 < \operatorname{Re} z < \pi/4$ onto the unit disk $|w| < 1$, with the four quarters of S mapped as indicated in Fig. 394. This mapping is conformal and one-to-one.

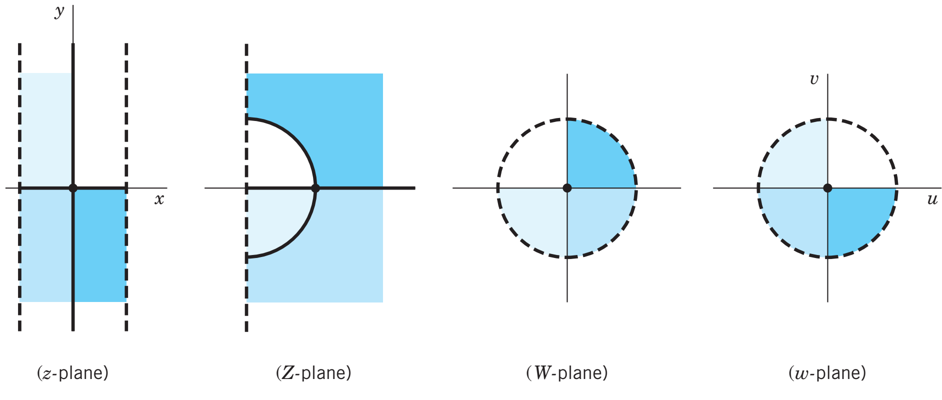


Fig. 394. Mapping by $w = \tan z$

PROBLEM SET 17.4

CONFORMAL MAPPING $w = e^z$

1. Find the image of $x = c = \text{const}$, $-\pi < y \leqq \pi$, under $w = e^z$.
2. Find the image of $y = k = \text{const}$, $-\infty < x \leqq \infty$, under $w = e^z$.

3–7 Find and sketch the image of the given region under $w = e^z$.

3. $-\frac{1}{2} \leqq x \leqq \frac{1}{2}, \quad -\pi \leqq y \leqq \pi$
4. $0 < x < 1, \quad \frac{1}{2} < y < 1$
5. $-\infty < x < \infty, \quad 0 \leqq y \leqq 2\pi$
6. $0 < x < \infty, \quad 0 < y < \pi/2$
7. $0 < x < 1, \quad 0 < y < \pi$
8. **CAS EXPERIMENT. Conformal Mapping.** If your CAS can do conformal mapping, use it to solve Prob. 7. Then increase y beyond π, say, to 50π or 100π. State what you expected. See what you get as the image. Explain.

CONFORMAL MAPPING $w = \sin z$

9. Find the points at which $w = \sin z$ is not conformal.
10. Sketch or graph the images of the lines $x = 0, \pm\pi/6, \pm\pi/3, \pm\pi/2$ under the mapping $w = \sin z$.

11–14 Find and sketch or graph the image of the given region under $w = \sin z$.

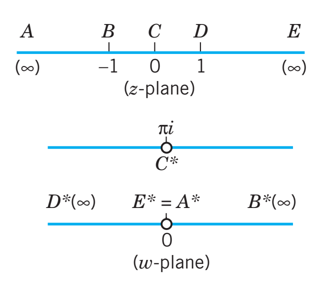
11. $0 < x < \pi/2, \quad 0 < y < 2$
12. $-\pi/4 < x < \pi/4, \quad 0 < y < 1$
13. $0 < x < 2\pi, \quad 1 < y < 3$
14. $0 < x < \pi/6, \quad -\infty < y < \infty$
15. Describe the mapping $w = \cosh z$ in terms of the mapping $w = \sin z$ and rotations and translations.
16. Find all points at which the mapping $w = \cosh 2\pi z$ is not conformal.
17. Find an analytic function that maps the region R bounded by the positive x- and y-semi-axes and the hyperbola $xy = \pi$ in the first quadrant onto the upper half-plane. *Hint.* First map R onto a horizontal strip.

CONFORMAL MAPPING $w = \cos z$

18. Find the images of the lines $y = k = \text{const}$ under the mapping $w = \cos z$.
19. Find the images of the lines $x = c = \text{const}$ under the mapping $w = \cos z$.

20–23 Find and sketch or graph the image of the given region under the mapping $w = \cos z$.

20. $0 < x < 2\pi, \quad \frac{1}{2} < y < 1$
21. $0 < x < \pi/2, \quad 0 < y < 2$ directly and from Prob. 11
22. $-1 < x < 1, \quad 0 \leqq y \leqq 1$
23. $\pi < x < 2\pi, \quad y < 0$
24. Find and sketch the image of the region $2 \leqq |z| \leqq 3$, $\pi/4 \leqq \theta \leqq \pi/2$ under the mapping $w = \text{Ln } z$.
25. Show that $w = \text{Ln} \dfrac{z-1}{z+1}$ maps the upper half-plane onto the horizontal strip $0 \leqq \text{Im } w \leqq \pi$ as shown in the figure.



Problem 25

17.5 Riemann Surfaces. *Optional*

One of the simplest but most ingenious ideas in complex analysis is that of **Riemann surfaces**. They allow multivalued relations, such as $w = \sqrt{z}$ or $w = \ln z$, to become single-valued and therefore functions in the usual sense. This works because the Riemann surfaces consist of several sheets that are connected at certain points (called branch points). Thus $w = \sqrt{z}$ will need two sheets, being single-valued on each sheet. How many sheets do you think $w = \ln z$ needs? Can you guess, by recalling Sec. 13.7? (The answer will be given at the end of this section). Let us start our systematic discussion.

The mapping given by

$$w = u + iv = z^2 \qquad (1) \qquad \text{(Sec. 17.1)}$$

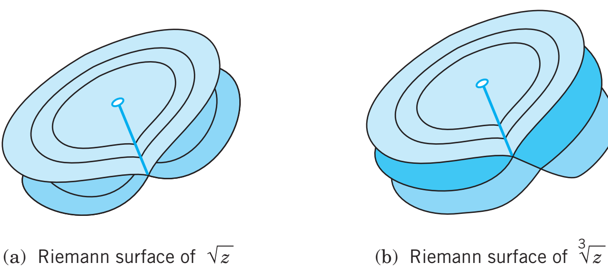
is conformal, except at $z = 0$, where $w' = 2z = 0$, At $z = 0$, angles are doubled under the mapping. Thus the right z-half-plane (including the positive y-axis) is mapped onto the full w-plane, cut along the negative half of the u-axis; this mapping is one-to-one. Similarly for the left z-half-plane (including the negative y-axis). Hence the image of the full z-plane under $w = z^2$ "covers the w-plane twice" in the sense that every $w \neq 0$ is the image of two z-points; if z_1 is one, the other is $-z_1$. For example, $z = i$ and $-i$ are both mapped onto $w = -1$.

Now comes the crucial idea. We place those two copies of the cut w-plane upon each other so that the upper sheet is the image of the right half z-plane R and the lower sheet is the image of the left half z-plane L. We join the two sheets crosswise along the cuts (along the negative u-axis) so that if z moves from R to L, its image can move from the upper to the lower sheet. The two origins are fastened together because $w = 0$ is the image of just one z-point, $z = 0$. The surface obtained is called a **Riemann surface** (Fig. 395a). $w = 0$ is called a "winding point" or **branch point**. $w = z^2$ maps the full z-plane onto this surface in a one-to-one manner.

By interchanging the roles of the variables z and w it follows that the double-valued relation

$$w = \sqrt{z} \qquad \text{(Sec. 13.2)} \tag{2}$$

becomes single-valued on the Riemann surface in Fig. 395a, that is, a function in the usual sense. We can let the upper sheet correspond to the principal value of $\sqrt{z}$. Its image is the right w-half-plane. The other sheet is then mapped onto the left w-half-plane.



(a) Riemann surface of $\sqrt{z}$ (b) Riemann surface of $\sqrt[3]{z}$

Fig. 395. Riemann surfaces

Similarly, the triple-valued relation $w = \sqrt[3]{z}$ becomes single-valued on the three-sheeted Riemann surface in Fig. 395b, which also has a branch point at $z = 0$.

The infinitely many-valued natural logarithm (Sec. 13.7)

$$w = \ln z = \text{Ln}\, z + 2n\pi i \qquad (n = 0, \pm 1, \pm 2, \cdots)$$

becomes single-valued on a Riemann surface consisting of infinitely many sheets, $w = \text{Ln}\, z$ corresponds to one of them. This sheet is cut along the negative x-axis and the upper edge of the slit is joined to the lower edge of the next sheet, which corresponds to the argument $\pi < \theta \leqq 3\pi$, that is, to

$$w = \text{Ln}\, z + 2\pi i.$$

The principal value $\text{Ln}\, z$ maps its sheet onto the horizontal strip $-\pi < v \leqq \pi$. The function $w = \text{Ln}\, z + 2\pi i$ maps its sheet onto the neighboring strip $\pi < v \leqq 3\pi$, and so on. The mapping of the points $z \neq 0$ of the Riemann surface onto the points of the w-plane is one-to-one. See also Example 5 in Sec. 17.1.

PROBLEM SET 17.5

1. If z moves from $z = \frac{1}{4}$ twice around the circle $|z| = \frac{1}{4}$, what does $w = \sqrt{z}$ do?

2. Show that the Riemann surface of $w = \sqrt{(z-1)(z-2)}$ has branch points at 1 and 2 sheets, which we may cut and join crosswise from 1 to 2. *Hint.* Introduce polar coordinates $z - 1 = r_1 e^{i\theta_1}$ and $z - 2 = r_2 e^{i\theta_2}$, so that $w = \sqrt{r_1 r_2}\, e^{i(\theta_1+\theta_2)/2}$.

3. Make a sketch, similar to Fig. 395, of the Riemann surface of $w = \sqrt[4]{z+1}$.

4–10 RIEMANN SURFACES

Find the branch points and the number of sheets of the Riemann surface.

4. $\sqrt{iz - 2 + i}$

5. $z^2 + \sqrt[3]{4z + i}$

6. $\ln(6z - 2i)$

7. $\sqrt[n]{z - z_0}$

8. $e^{\sqrt{z}}$, $\sqrt{e^z}$

9. $\sqrt{z^3 + z}$

10. $\sqrt{(4 - z^2)(1 - z^2)}$

CHAPTER 17 REVIEW QUESTIONS AND PROBLEMS

1. What is a conformal mapping? Why does it occur in complex analysis?

2. At what points are $w = z^5 - z$ and $w = \cos(\pi z^2)$ not conformal?

3. What happens to angles at z_0 under a mapping $w = f(z)$ if $f'(z_0) = 0$, $f''(z_0) = 0$, $f'''(z_0) \neq 0$?

4. What is a linear fractional transformation? What can you do with it? List special cases.

5. What is the extended complex plane? Ways of introducing it?

6. What is a fixed point of a mapping? Its role in this chapter? Give examples.

7. How would you find the image of $x = \text{Re}\, z = 1$ under $w = iz, z^2, e^z, 1/z$?

8. Can you remember mapping properties of $w = \ln z$?

9. What mapping gave the Joukowski airfoil? Explain details.

10. What is a Riemann surface? Its motivation? Its simplest example.

11–16 MAPPING $w = z^2$

Find and sketch the image of the given region or curve under $w = z^2$.

11. $1 < |z| < 2$, $\quad |\arg z| < \pi/8$

12. $1/\sqrt{\pi} < |z| < \sqrt{\pi}$, $\quad 0 < \arg z < \pi/2$

13. $-4 < xy < 4$

14. $0 < y < 2$

15. $x = -1, 1$

16. $y = -2, 2$

17–22 MAPPING $w = 1/z$

Find and sketch the image of the given region or curve under $w = 1/z$.

17. $|z| < 1$

18. $|z| < 1$, $\quad 0 < \arg z < \pi/2$

19. $2 < |z| < 3$, $\quad y > 0$

20. $0 \leqq \arg z \leqq \pi/4$

21. $(x - \frac{1}{2})^2 + y^2 = \frac{1}{4}$, $\quad y > 0$

22. $z = 1 + iy \quad (-\infty < y < \infty)$

23–28 LINEAR FRACTIONAL TRANSFORMATIONS (LFTs)

Find the LFT that maps

23. $-1, 0, 1$ onto $4 + 3i, 5i/2, 4 - 3i$, respectively

24. $0, 2, 4$ onto $\infty, \frac{1}{2}, \frac{1}{4}$, respectively

25. $1, i, -i$ onto $i, -1, 1$, respectively

26. $0, 1, 2$ onto $2i, 1 + 2i, 2 + 2i$, respectively

27. $0, 1, \infty$ onto $\infty, 1, 0$, respectively

28. $-1, -i, i$ onto $1 - i, 2, 0$, respectively

29–34 FIXED POINTS

Find the fixed points of the mapping

29. $w = (2 + i)z$

30. $w = z^4 + z - 64$

31. $w = (3z + 2)/(z - 1)$

32. $(2iz - 1)/(z + 2i)$

33. $w = z^5 + 10z^3 + 10z$

34. $w = (iz + 5)/(5z + i)$

35–40 GIVEN REGIONS

Find an analytic function $w = f(z)$ that maps

35. The infinite strip $0 < y < \pi/4$ onto the upper half-plane $v > 0$.

36. The quarter-disk $|z| < 1, x > 0, y > 0$ onto the exterior of the unit circle $|w| = 1$.

37. The sector $0 < \arg z < \pi/2$ onto the region $u < 1$.

38. The interior of the unit circle $|z| = 1$ onto the exterior of the circle $|w + 2| = 2$.

39. The region $x > 0, y > 0, xy < c$ onto the strip $0 < v < 1$.

40. The semi-disk $|z| < 2, y > 0$ onto the exterior of the circle $|w - \pi| = \pi$.

SUMMARY OF CHAPTER 17
Conformal Mapping

A complex function $w = f(z)$ gives a **mapping** of its domain of definition in the complex z-plane onto its range of values in the complex w-plane. If $f(z)$ is *analytic*, this mapping is **conformal**, that is, angle-preserving: the images of any two intersecting curves make the same angle of intersection, in both magnitude and sense, as the curves themselves (Sec. 17.1). Exceptions are the points at which $f'(z) = 0$ ("**critical points**," e.g. $z = 0$ for $w = z^2$).

For mapping properties of e^z, $\cos z$, $\sin z$ etc. see Secs. 17.1 and 17.4.

Linear fractional transformations, also called *Möbius transformations*

(1) $$w = \frac{az + b}{cz + d}$$ (Secs. 17.2, 17.3)

($ad - bc \neq 0$) map the extended complex plane (Sec. 17.2) onto itself. They solve the problems of mapping half-planes onto half-planes or disks, and disks onto disks or half-planes. Prescribing the images of three points determines (1) uniquely.

Riemann surfaces (Sec. 17.5) consist of several sheets connected at certain points called *branch points.* On them, multivalued relations become single-valued, that is, functions in the usual sense. *Examples.* For $w = \sqrt{z}$ we need two sheets (with branch point 0) since this relation is doubly-valued. For $w = \ln z$ we need infinitely many sheets since this relation is infinitely many-valued (see Sec. 13.7).

Complex Analysis and Potential Theory

In Chapter 17 we developed the *geometric approach of conformal mapping*. This meant that, for a complex analytic function $w = f(z)$ defined in a domain D of the z-plane, we associated with each point in D a corresponding point in the w-plane. This gave us a *conformal mapping* (angle-preserving), except at critical points where $f'(z) = 0$.

Now, in this chapter, we shall apply conformal mappings to potential problems. This will lead to boundary value problems and many engineering applications in electrostatics, heat flow, and fluid flow. More details are as follows.

Recall that Laplace's equation $\nabla^2\Phi = 0$ is one of the most important PDEs in engineering mathematics because it occurs in gravitation (Secs. 9.7, 12.11), electrostatics (Sec. 9.7), steady-state heat conduction (Sec. 12.5), incompressible fluid flow, and other areas. The theory of this equation is called **potential theory** (although "potential" is also used in a more general sense in connection with gradients (see Sec. 9.7)). Because we want to treat this equation with complex analytic methods, we restrict our discussion to the "two-dimensional case." Then Φ depends only on two Cartesian coordinates x and y, and Laplace's equation becomes

$$\nabla^2\Phi = \Phi_{xx} + \Phi_{yy} = 0.$$

An important idea then is that its solutions Φ are closely related to complex analytic functions $\Phi + i\Psi$ as shown in Sec. 13.4. (*Remark:* We use the notation $\Phi + i\Psi$ to free u and v, which will be needed in conformal mapping $u + iv$.) This important relation is the main reason for using complex analysis in problems of physics and engineering.

We shall examine this connection between Laplace's equation and complex analytic functions and illustrate it by modeling applications from electrostatics (Secs. 18.1, 18.2), heat conduction (Sec. 18.3), and hydrodynamics (Sec. 18.4). This in turn will lead to ***boundary value problems*** in two-dimensional potential theory. As a result, some of the functions of Chap. 17 will be used to transform complicated regions into simpler ones.

Section 18.5 will derive the important Poisson formula for potentials in a circular disk. Section 18.6 will deal with **harmonic functions**, which, as you recall, are solutions of Laplace's equation and have continuous second partial derivatives. In that section we will show how results on analytic functions can be used to characterize properties of harmonic functions.

Prerequisite: Chaps. 13, 14, 17.
References and Answers to Problems: App. 1 Part D, App. 2.

18.1 Electrostatic Fields

The electrical force of attraction or repulsion between charged particles is governed by Coulomb's law (see Sec. 9.7). This force is the gradient of a function Φ, called the **electrostatic potential**. At any points free of charges, Φ is a solution of Laplace's equation

$$\nabla^2 \Phi = 0.$$

The surfaces $\Phi = \text{const}$ are called **equipotential surfaces**. At each point P at which the gradient of Φ is not the zero vector, it is perpendicular to the surface $\Phi = \text{const}$ through P; that is, the electrical force has the direction perpendicular to the equipotential surface. (See also Secs. 9.7 and 12.11.)

The problems we shall discuss in this entire chapter are **two-dimensional** (for the reason just given in the chapter opening), that is, they model physical systems that lie in three-dimensional space (of course!), but are such that the potential Φ is independent of one of the space coordinates, so that Φ depends only on two coordinates, which we call x and y. Then **Laplace's equation** becomes

$$\nabla^2 \Phi = \frac{\partial^2 \Phi}{\partial x^2} + \frac{\partial^2 \Phi}{\partial y^2} = 0. \tag{1}$$

Fig. 396. Potential in Example 1

Equipotential surfaces now appear as **equipotential lines** (curves) in the xy-plane.

Let us illustrate these ideas by a few simple examples.

EXAMPLE 1 Potential Between Parallel Plates

Find the potential Φ of the field between two parallel conducting plates extending to infinity (Fig. 396), which are kept at potentials Φ_1 and Φ_2, respectively.

Solution. From the shape of the plates it follows that Φ depends only on x, and Laplace's equation becomes $\Phi'' = 0$. By integrating twice we obtain $\Phi = ax + b$, where the constants a and b are determined by the given boundary values of Φ on the plates. For example, if the plates correspond to $x = -1$ and $x = 1$, the solution is

$$\Phi(x) = \tfrac{1}{2}(\Phi_2 - \Phi_1)x + \tfrac{1}{2}(\Phi_2 + \Phi_1).$$

The equipotential surfaces are parallel planes. ■

EXAMPLE 2 Potential Between Coaxial Cylinders

Find the potential Φ between two coaxial conducting cylinders extending to infinity on both ends (Fig. 397) and kept at potentials Φ_1 and Φ_2, respectively.

Solution. Here Φ depends only on $r = \sqrt{x^2 + y^2}$, for reasons of symmetry, and Laplace's equation $r^2 u_{rr} + r u_r + u_{\theta\theta} = 0$ [(5), Sec. 12.10] with $u_{\theta\theta} = 0$ and $u = \Phi$ becomes $r\Phi'' + \Phi' = 0$. By separating variables and integrating we obtain

$$\frac{\Phi''}{\Phi'} = -\frac{1}{r}, \qquad \ln \Phi' = -\ln r + \tilde{a}, \qquad \Phi' = \frac{a}{r}, \qquad \Phi = a \ln r + b$$

and a and b are determined by the given values of Φ on the cylinders. Although no infinitely extended conductors exist, the field in our idealized conductor will approximate the field in a long finite conductor in that part which is far away from the two ends of the cylinders. ■

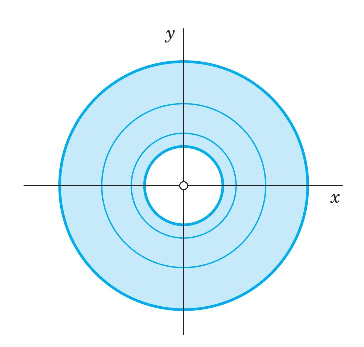


Fig. 397. Potential in Example 2

EXAMPLE 3 **Potential in an Angular Region**

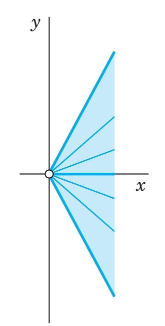


Fig. 398. Potential in Example 3

Find the potential Φ between the conducting plates in Fig. 398, which are kept at potentials Φ_1 (the lower plate) and Φ_2, and make an angle α, where $0 < \alpha \leqq \pi$. (In the figure we have $\alpha = 120° = 2\pi/3$.)

Solution. $\theta = \text{Arg}\, z\ (z = x + iy \neq 0)$ is constant on rays $\theta = \text{const}$. It is harmonic since it is the imaginary part of an analytic function, $\text{Ln}\, z$ (Sec. 13.7). Hence the solution is

$$\Phi(x, y) = a + b\, \text{Arg}\, z$$

with a and b determined from the two boundary conditions (given values on the plates)

$$a + b(-\tfrac{1}{2}\alpha) = \Phi_1, \qquad a + b(\tfrac{1}{2}\alpha) = \Phi_2.$$

Thus $a = (\Phi_2 + \Phi_1)/2$, $b = (\Phi_2 - \Phi_1)/\alpha$. The answer is

$$\Phi(x, y) = \frac{1}{2}(\Phi_2 + \Phi_1) + \frac{1}{\alpha}(\Phi_2 - \Phi_1)\,\theta, \qquad \theta = \arctan\frac{y}{x}.$$ ■

Complex Potential

Let $\Phi(x, y)$ be harmonic in some domain D and $\Psi(x, y)$ a harmonic conjugate of Φ in D. (Note the change of notation from u and v of Sec. 13.4 to Φ and Ψ. From the next section on, we had to free u and v for use in conformal mapping. Then

$$F(z) = \Phi(x, y) + i\Psi(x, y) \tag{2}$$

is an analytic function of $z = x + iy$. This function F is called the **complex potential** corresponding to the real potential Φ. Recall from Sec. 13.4 that for given Φ, a conjugate Ψ is uniquely determined except for an additive real constant. Hence we may say ***the*** complex potential, without causing misunderstandings.

The use of F has two advantages, a technical one and a physical one. Technically, F is easier to handle than real or imaginary parts, in connection with methods of complex analysis. Physically, Ψ has a meaning. By conformality, the curves $\Psi = \text{const}$ intersect the equipotential lines $\Phi = \text{const}$ in the xy-plane at right angles [except where $F'(z) = 0$]. Hence they have the direction of the electrical force and, therefore, are called **lines of force**. They are the paths of moving charged particles (electrons in an electron microscope, etc.).

EXAMPLE 4 Complex Potential

In Example 1, a conjugate is $\Psi = ay$. It follows that the complex potential is

$$F(z) = az + b = ax + b + iay,$$

and the lines of force are horizontal straight lines $y = \text{const}$ parallel to the x-axis. ■

EXAMPLE 5 Complex Potential

In Example 2 we have $\Phi = a \ln r + b = a \ln |z| + b$. A conjugate is $\Psi = a \operatorname{Arg} z$. Hence the complex potential is

$$F(z) = a \operatorname{Ln} z + b$$

and the lines of force are straight lines through the origin. $F(z)$ may also be interpreted as the complex potential of a source line (a wire perpendicular to the xy-plane) whose trace in the xy-plane is the origin. ■

EXAMPLE 6 Complex Potential

In Example 3 we get $F(z)$ by noting that $i \operatorname{Ln} z = i \ln |z| - \operatorname{Arg} z$, multiplying this by $-b$, and adding a:

$$F(z) = a - ib \operatorname{Ln} z = a + b \operatorname{Arg} z - ib \ln |z|.$$

We see from this that the lines of force are concentric circles $|z| = \text{const}$. Can you sketch them? ■

Superposition

More complicated potentials can often be obtained by superposition.

EXAMPLE 7 Potential of a Pair of Source Lines (a Pair of Charged Wires)

Determine the potential of a pair of oppositely charged source lines of the same strength at the points $z = c$ and $z = -c$ on the real axis.

Solution. From Examples 2 and 5 it follows that the potential of each of the source lines is

$$\Phi_1 = K \ln |z - c| \qquad \text{and} \qquad \Phi_2 = -K \ln |z + c|,$$

respectively. Here the real constant K measures the strength (amount of charge). These are the real parts of the complex potentials

$$F_1(z) = K \operatorname{Ln}(z - c) \qquad \text{and} \qquad F_2(z) = -K \operatorname{Ln}(z + c).$$

Hence the complex potential of the combination of the two source lines is

$$F(z) = F_1(z) + F_2(z) = K[\operatorname{Ln}(z - c) - \operatorname{Ln}(z + c)]. \tag{3}$$

The **equipotential lines** are the curves

$$\Phi = \operatorname{Re} F(z) = K \ln \left| \frac{z - c}{z + c} \right| = \text{const}, \qquad \text{thus} \qquad \left| \frac{z - c}{z + c} \right| = \text{const}.$$

These are circles, as you may show by direct calculation. The **lines of force** are

$$\Psi = \operatorname{Im} F(z) = K[\operatorname{Arg}(z - c) - \operatorname{Arg}(z + c)] = \text{const}.$$

We write this briefly (Fig. 399)

$$\Psi = K(\theta_1 - \theta_2) = \text{const}.$$

Now $\theta_1 - \theta_2$ is the angle between the line segments from z to c and $-c$ (Fig. 399). Hence the lines of force are the curves along each of which the line segment S: $-c \leqq x \leqq c$ appears under a constant angle. These curves are the totality of circular arcs over S, as is (or should be) known from elementary geometry. Hence the lines of force are circles. Figure 400 shows some of them together with some equipotential lines.

In addition to the interpretation as the potential of two source lines, this potential could also be thought of as the potential between two circular cylinders whose axes are parallel but do not coincide, or as the potential between two equal cylinders that lie outside each other, or as the potential between a cylinder and a plane wall. Explain this using Fig. 400. ■

The idea of the complex potential as just explained is the key to a close relation of potential theory to complex analysis and will recur in heat flow and fluid flow.

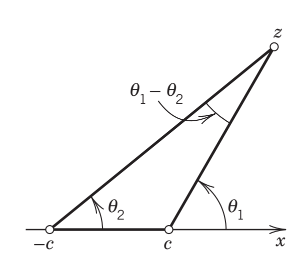


Fig. 399. Arguments in Example 7

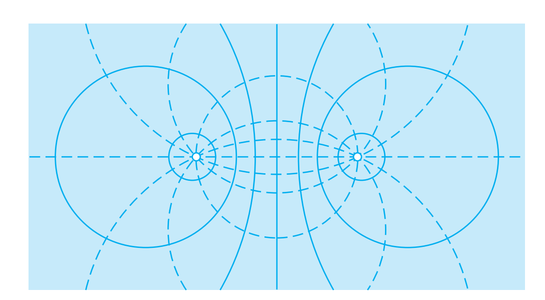


Fig. 400. Equipotential lines and lines of force (dashed) in Example 7

PROBLEM SET 18.1

1–4 COAXIAL CYLINDERS

Find and sketch the potential between two coaxial cylinders of radii r_1 and r_2 having potential U_1 and U_2, respectively.

1. $r_1 = 2.5$ mm, $r_2 = 4.0$ cm, $U_1 = 0$ V, $U_2 = 220$ V
2. $r_1 = 1$ cm, $r_2 = 2$ cm, $U_1 = 400$ V, $U_2 = 0$ V
3. $r_1 = 10$ cm, $r_2 = 1$ m, $U_1 = 10$ kV, $U_2 = -10$ kV
4. If $r_1 = 2$ cm, $r_2 = 6$ cm and $U_1 = 300$ V, $U_2 = 100$ V, respectively, is the potential at $r = 4$ cm equal to 200 V? Less? More? Answer without calculation. Then calculate and explain.

5–7 PARALLEL PLATES

Find and sketch the potential between the parallel plates having potentials U_1 and U_2. Find the complex potential.

5. Plates at $x_1 = -5$ cm, $x_2 = 5$ cm, potentials $U_1 = 250$ V, $U_2 = 500$ V, respectively.
6. Plates at $y = x$ and $y = x + k$, potentials $U_1 = 0$ V, $U_2 = 220$ V, respectively.
7. Plates at $x_1 = 12$ cm, $x_2 = 24$ cm, potentials $U_1 = 20$ kV, $U_2 = 8$ kV, respectively.
8. **CAS EXPERIMENT. Complex Potentials.** Graph the equipotential lines and lines of force in (a)–(d) (four graphs, Re $F(z)$ and Im $F(z)$ on the same axes). Then explore further complex potentials of your choice with the purpose of discovering configurations that might be of practical interest.

   (a) $F(z) = z^2$ **(b)** $F(z) = iz^2$
   (c) $F(z) = 1/z$ **(d)** $F(z) = i/z$

9. **Argument.** Show that $\Phi = \theta/\pi = (1/\pi) \arctan(y/x)$ is harmonic in the upper half-plane and satisfies the boundary condition $\Phi(x, 0) = 1$ if $x < 0$ and 0 if $x > 0$, and the corresponding complex potential is $F(z) = -(i/\pi) \operatorname{Ln} z$.
10. **Conformal mapping.** Map the upper z-half-plane onto $|w| \leqq 1$ so that $0, \infty, -1$ are mapped onto $1, i, -i$, respectively. What are the boundary conditions on $|w| = 1$ resulting from the potential in Prob. 9? What is the potential at $w = 0$?
11. **Text Example 7.** Verify, by calculation, that the equipotential lines are circles.

12–15 OTHER CONFIGURATIONS

12. Find and sketch the potential between the axes (potential 500 V) and the hyperbola $xy = 4$ (potential 100 V).

13. Arccos. Show that $F(z) = \arccos z$ (defined in Problem Set 13.7) gives the potential of a slit in Fig. 401.

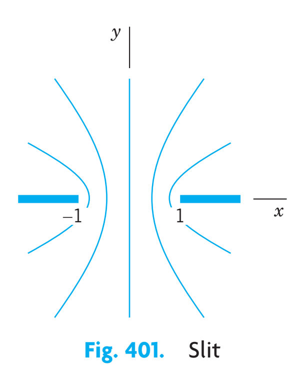


Fig. 401. Slit

14. Arccos. Show that $F(z)$ in Prob. 13 gives the potentials in Fig. 402.

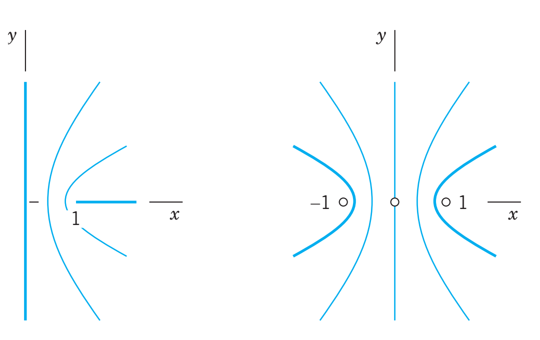


Fig. 402. Other apertures

15. Sector. Find the real and complex potentials in the sector $-\pi/6 \leqq \theta \leqq \pi/6$ between the boundary $\theta = \pm\pi/6$, kept at 0 V, and the curve $x^3 - 3xy^2 = 1$, kept at 220 V.

18.2 Use of Conformal Mapping. Modeling

We have just explored the close relation between potential theory and complex analysis. This relationship is so close because complex potentials can be modeled in complex analysis. In this section we shall explore the close relation that results from the use of conformal mapping in modeling and solving ***boundary value problems*** for the Laplace equation. The process consists of finding a solution of the equation in some domain, assuming given values on the boundary (***Dirichlet problem***, see also Sec. 12.6). The key idea is then to use conformal mapping to map a given domain onto one for which the solution is known or can be found more easily. This solution thus obtained is then mapped back to the given domain. The reason this approach works is due to Theorem 1, which asserts that harmonic functions remain harmonic under conformal mapping:

THEOREM 1

Harmonic Functions Under Conformal Mapping

Let Φ^* be harmonic in a domain D^* in the w-plane. Suppose that $w = u + iv = f(z)$ is analytic in a domain D in the z-plane and maps D conformally onto D^*. Then the function

$$\Phi(x, y) = \Phi^*(u(x, y), v(x, y)) \tag{1}$$

is harmonic in D.

PROOF The composite of analytic functions is analytic, as follows from the chain rule. Hence, taking a harmonic conjugate $\Psi^*(u, v)$ of Φ^*, as defined in Sec. 13.4, and forming the analytic function $F^*(w) = \Phi^*(u, v) + i\Psi^*(u, v)$ we conclude that $F(z) = F^*(f(z))$ is analytic in D. Hence its real part $\Phi(x, y) = \operatorname{Re} F(z)$ is harmonic in D. This completes the proof.

We mention without proof that if D^* is simply connected (Sec. 14.2), then a harmonic conjugate of Φ^* exists. Another proof of Theorem 1 without the use of a harmonic conjugate is given in App. 4. ■

EXAMPLE 1 **Potential Between Noncoaxial Cylinders**

Model the electrostatic potential between the cylinders C_1: $|z| = 1$ and C_2: $|z - \frac{2}{5}| = \frac{2}{5}$ in Fig. 403. Then give the solution for the case that C_1 is grounded, $U_1 = 0$ V, and C_2 has the potential $U_2 = 110$ V.

Solution. We map the unit disk $|z| = 1$ onto the unit disk $|w| = 1$ in such a way that C_2 is mapped onto some cylinder C_2^*: $|w| = r_0$. By (3), Sec. 17.3, a linear fractional transformation mapping the unit disk onto the unit disk is

$$w = \frac{z - b}{bz - 1} \tag{2}$$

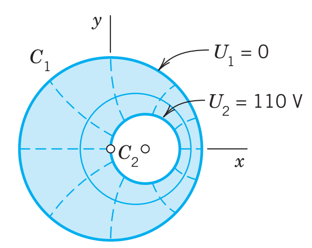


Fig. 403. Example 1: z-plane

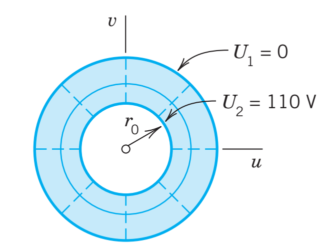


Fig. 404. Example 1: w-plane

where we have chosen $b = z_0$ real without restriction. z_0 is of no immediate help here because centers of circles do not map onto centers of the images, in general. However, we now have two free constants b and r_0 and shall succeed by imposing two reasonable conditions, namely, that 0 and $\frac{4}{5}$ (Fig. 403) should be mapped onto r_0 and $-r_0$ (Fig. 404), respectively. This gives by (2)

$$r_0 = \frac{0 - b}{0 - 1} = b, \qquad \text{and with this,} \qquad -r_0 = \frac{\frac{4}{5} - b}{4b/5 - 1} = \frac{\frac{4}{5} - r_0}{4r_0/5 - 1},$$

a quadratic equation in r_0 with solutions $r_0 = 2$ (no good because $r_0 < 1$) and $r_0 = \frac{1}{2}$. Hence our mapping function (2) with $b = \frac{1}{2}$ becomes that in Example 5 of Sec. 17.3,

$$w = f(z) = \frac{2z - 1}{z - 2}. \tag{3}$$

From Example 5 in Sec. 18.1, writing w for z we have as the complex potential in the w-plane the function $F^*(w) = a \operatorname{Ln} w + k$ and from this the real potential

$$\Phi^*(u, v) = \operatorname{Re} F^*(w) = a \ln |w| + k.$$

This is our model. We now determine a and k from the boundary conditions. If $|w| = 1$, then $\Phi^* = a \ln 1 + k = 0$, hence $k = 0$. If $|w| = r_0 = \frac{1}{2}$, then $\Phi^* = a \ln (\frac{1}{2}) = 110$, hence $a = 110/\ln (\frac{1}{2}) = -158.7$. Substitution of (3) now gives the desired solution in the given domain in the z-plane

$$F(z) = F^*(f(z)) = a \operatorname{Ln} \frac{2z - 1}{z - 2}.$$

The real potential is

$$\Phi(x, y) = \operatorname{Re} F(z) = a \ln \left| \frac{2z - 1}{z - 2} \right|, \qquad a = -158.7.$$

Can we "see" this result? Well, $\Phi(x, y) = \text{const}$ if and only if $|(2z - 1)/(z - 2)| = \text{const}$, that is, $|w| = \text{const}$ by (2) with $b = \frac{1}{2}$. These circles are images of circles in the z-plane because the inverse of a linear fractional transformation is linear fractional (see (4), Sec. 17.2), and any such mapping maps circles onto circles (or straight lines), by Theorem 1 in Sec. 17.2. Similarly for the rays $\arg w = \text{const}$. Hence the equipotential lines $\Phi(x, y) = \text{const}$ are circles, and the lines of force are circular arcs (dashed in Fig. 404). These two families of curves intersect orthogonally, that is, at right angles, as shown in Fig. 404. ■

EXAMPLE 2 Potential Between Two Semicircular Plates

Model the potential between two semicircular plates P_1 and P_2 in Fig. 405 having potentials -3000 V and 3000 V, respectively. Use Example 3 in Sec. 18.1 and conformal mapping.

Solution. ***Step 1.*** We map the unit disk in Fig. 405 onto the right half of the w-plane (Fig. 406) by using the linear fractional transformation in Example 3, Sec. 17.3:

$$w = f(z) = \frac{1+z}{1-z}.$$

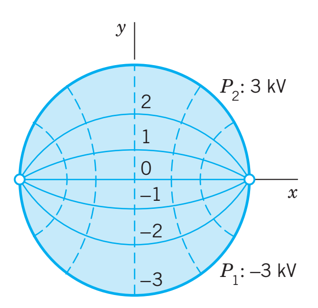


Fig. 405. Example 2: z-plane

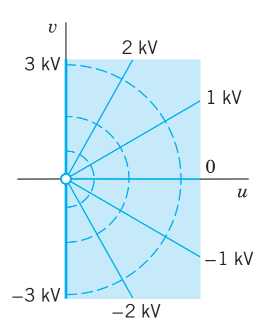


Fig. 406. Example 2: w-plane

The boundary $|z| = 1$ is mapped onto the boundary $u = 0$ (the v-axis), with $z = -1, i, 1$ going onto $w = 0, i, \infty$, respectively, and $z = -i$ onto $w = -i$. Hence the upper semicircle of $|z| = 1$ is mapped onto the upper half, and the lower semicircle onto the lower half of the v-axis, so that the boundary conditions in the w-plane are as indicated in Fig. 406.

Step 2. We determine the potential $\Phi^*(u, v)$ in the right half-plane of the w-plane. Example 3 in Sec. 18.1 with $\alpha = \pi$, $U_1 = -3000$, and $U_2 = 3000$ [with $\Phi^*(u, v)$ instead of $\Phi(x, y)$] yields

$$\Phi^*(u, v) = \frac{6000}{\pi}\varphi, \qquad \varphi = \arctan\frac{v}{u}.$$

On the positive half of the imaginary axis ($\varphi = \pi/2$), this equals 3000 and on the negative half -3000, as it should be. Φ^* is the real part of the complex potential

$$F^*(w) = -\frac{6000\,i}{\pi}\operatorname{Ln} w.$$

Step 3. We substitute the mapping function into F^* to get the complex potential $F(z)$ in Fig. 405 in the form

$$F(z) = F^*(f(z)) = -\frac{6000\,i}{\pi}\operatorname{Ln}\frac{1+z}{1-z}.$$

The real part of this is the potential we wanted to determine:

$$\Phi(x, y) = \operatorname{Re} F(z) = \frac{6000}{\pi}\operatorname{Im}\operatorname{Ln}\frac{1+z}{1-z} = \frac{6000}{\pi}\operatorname{Arg}\frac{1+z}{1-z}.$$

As in Example 1 we conclude that the equipotential lines $\Phi(x, y) = \text{const}$ are circular arcs because they correspond to $\operatorname{Arg}[(1+z)/(1-z)] = \text{const}$, hence to $\operatorname{Arg} w = \text{const}$. Also, $\operatorname{Arg} w = \text{const}$ are rays from 0 to ∞, the images of $z = -1$ and $z = 1$, respectively. Hence the equipotential lines all have -1 and 1 (the points where the boundary potential jumps) as their endpoints (Fig. 405). The lines of force are circular arcs, too, and since they must be orthogonal to the equipotential lines, their centers can be obtained as intersections of tangents to the unit circle with the x-axis, (Explain!) ■

Further examples can easily be constructed. Just take any mapping $w = f(z)$ in Chap. 17, a domain D in the z-plane, its image D^* in the w-plane, and a potential Φ^* in D^*. Then (1) gives a potential in D. Make up some examples of your own, involving, for instance, linear fractional transformations.

Basic Comment on Modeling

We formulated the examples in this section as models on the electrostatic potential. It is quite important to realize that this is accidental. We could equally well have phrased everything in terms of (time-independent) heat flow; then instead of voltages we would have had temperatures, the equipotential lines would have become isotherms (= lines of constant temperature), and the lines of the electrical force would have become lines along which heat flows from higher to lower temperatures (more on this in the next section). Or we could have talked about fluid flow; then the electrostatic lines of force would have become streamlines (more on this in Sec. 18.4). What we again see here is the ***unifying power of mathematics***: different phenomena and systems from different areas in physics having the same types of model can be treated by the same mathematical methods. What differs from area to area is just the kinds of problems that are of practical interest.

PROBLEM SET 18.2

1. **Derivation of (3) from (2).** Verify the steps.
2. **Second proof.** Give the details of the steps given on p. A93 of the book. What is the point of that proof?

3–5 APPLICATION OF THEOREM 1

3. Find the potential Φ in the region R in the first quadrant of the z-plane bounded by the axes (having potential U_1) and the hyperbola $y = 1/x$ (having potential U_2) by mapping R onto a suitable infinite strip. Show that Φ is harmonic. What are its boundary values?
4. Let $\Phi^* = 4uv$, $w = f(z) = e^z$, and D: $x < 0$, $0 < y < \pi$. Find Φ. What are its boundary values?
5. **CAS PROJECT. Graphing Potential Fields.** Graph equipotential lines **(a)** in Example 1 of the text, **(b)** if the complex potential is $F(z) = z^2, iz^2, e^z$. **(c)** Graph the equipotential surfaces for $F(z) = \text{Ln}\, z$ as cylinders in space.
6. Apply Theorem 1 to $\Phi^*(u, v) = u^2 - v^2$, $w = f(z) = e^z$, and any domain D, showing that the resulting potential Φ is harmonic.
7. **Rectangle, sin *z*.** Let D: $0 \leqq x \leqq \frac{1}{2}\pi$, $0 \leqq y \leqq 1$; D^* the image of D under $w = \sin z$; and $\Phi^* = u^2 - v^2$. What is the corresponding potential Φ in D? What are its boundary values? Sketch D and D^*.
8. **Conjugate potential.** What happens in Prob. 7 if you replace the potential by its conjugate harmonic?
9. **Translation.** What happens in Prob. 7 if we replace $\sin z$ by $\cos z = \sin\,(z + \frac{1}{2}\pi)$?
10. **Noncoaxial Cylinders.** Find the potential between the cylinders C_1: $|z| = 1$ (potential $U_1 = 0$) and C_2: $|z - c| = c$ (potential $U_2 = 220$ V), where $0 < c < \frac{1}{2}$. Sketch or graph equipotential lines and their orthogonal trajectories for $c = \frac{1}{4}$. Can you guess how the graph changes if you increase c $(< \frac{1}{2})$?
11. **On Example 2.** Verify the calculations.
12. Show that in Example 2 the y-axis is mapped onto the unit circle in the w-plane.
13. At $z = \pm 1$ in Fig. 405 the tangents to the equipotential lines as shown make equal angles. Why?
14. Figure 405 gives the impression that the potential on the y-axis changes more rapidly near 0 than near $\pm i$. Can you verify this?
15. **Angular region.** By applying a suitable conformal mapping, obtain from Fig. 406 the potential Φ in the sector $-\frac{1}{4}\pi < \text{Arg}\, z < \frac{1}{4}\pi$ such that $\Phi = -3$ kV if $\text{Arg}\, z = -\frac{1}{4}\pi$ and $\Phi = 3$ kV if $\text{Arg}\, z = \frac{1}{4}\pi$.
16. Solve Prob. 15 if the sector is $-\frac{1}{8}\pi < \text{Arg}\, z < \frac{1}{8}\pi$.
17. **Another extension of Example 2.** Find the linear fractional transformation $z = g(Z)$ that maps $|Z| \leqq 1$ onto $|z| \leqq 1$ with $Z = i/2$ being mapped onto $z = 0$. Show that $Z_1 = 0.6 + 0.8i$ is mapped onto $z = -1$ and $Z_2 = -0.6 + 0.8i$ onto $z = 1$, so that the equipotential lines of Example 2 look in $|Z| \leqq 1$ as shown in Fig. 407.

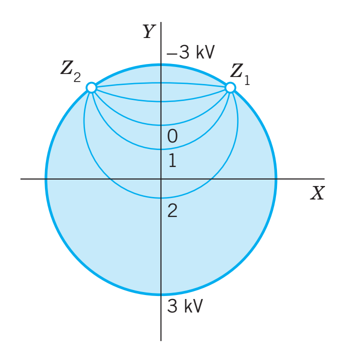


Fig. 407. Problem 17

18. The equipotential lines in Prob. 17 are circles. Why?

19. Jump on the boundary. Find the complex and real potentials in the upper half-plane with boundary values 5 kV if $x < 2$ and 0 if $x > 2$ on the x-axis.

20. Jumps. Do the same task as in Prob. 19 if the boundary values on the x-axis are V_0 when $-a < x < a$ and 0 elsewhere.

18.3 Heat Problems

Heat conduction in a body of homogeneous material is modeled by the heat equation

$$T_t = c^2 \nabla^2 T$$

where the function T is temperature, $T_t = \partial T/\partial t$, t is time, and c^2 is a positive constant (specific to the material of the body; see Sec. 12.6).

Now if a heat flow problem is **steady**, that is, independent of time, we have $T_t = 0$. If it is also two-dimensional, then the heat equation reduces to

$$\nabla^2 T = T_{xx} + T_{yy} = 0, \tag{1}$$

which is the two-dimensional Laplace equation. Thus we have shown that we can model a two-dimensional steady heat flow problem by *Laplace's equation*.

Furthermore we can treat this heat flow problem by methods of complex analysis, since T (or $T(x, y)$) is the real part of the **complex heat potential**

$$F(z) = T(x, y) + i\Psi(x, y).$$

We call $T(x, y)$ the heat potential. The curves $T(x, y) = \text{const}$ are called **isotherms**, which means lines of constant temperature. The curves $\Psi(x, y) = \text{const}$ are called **heat flow lines** because heat flows along them from higher temperatures to lower temperatures.

It follows that all the examples considered so far (Secs. 18.1, 18.2) can now be reinterpreted as problems on heat flow. The electrostatic equipotential lines $\Phi(x, y) = \text{const}$ now become isotherms $T(x, y) = \text{const}$, and the lines of electrical force become lines of heat flow, as in the following two problems.

EXAMPLE 1 Temperature Between Parallel Plates

Find the temperature between two parallel plates $x = 0$ and $x = d$ in Fig. 408 having temperatures 0 and 100°C, respectively.

Solution. As in Example 1 of Sec. 18.1 we conclude that $T(x, y) = ax + b$. From the boundary conditions, $b = 0$ and $a = 100/d$. The answer is

$$T(x, y) = \frac{100}{d} x \text{ [°C]}.$$

The corresponding complex potential is $F(z) = (100/d)z$. Heat flows horizontally in the negative x-direction along the lines $y = \text{const}$. ■

EXAMPLE 2 Temperature Distribution Between a Wire and a Cylinder

Find the temperature field around a long thin wire of radius $r_1 = 1$ mm that is electrically heated to $T_1 = 500$°F and is surrounded by a circular cylinder of radius $r_2 = 100$ mm, which is kept at temperature $T_2 = 60$°F by cooling it with air. See Fig. 409. (The wire is at the origin of the coordinate system.)

Solution. T depends only on r, for reasons of symmetry. Hence, as in Sec. 18.1 (Example 2),

$$T(x, y) = a \ln r + b.$$

The boundary conditions are

$$T_1 = 500 = a \ln 1 + b, \qquad T_2 = 60 = a \ln 100 + b.$$

Hence $b = 500$ (since $\ln 1 = 0$) and $a = (60 - b)/\ln 100 = -95.54$. The answer is

$$T(x, y) = 500 - 95.54 \ln r \,[°\text{F}].$$

The isotherms are concentric circles. Heat flows from the wire radially outward to the cylinder. Sketch T as a function of r. Does it look physically reasonable?

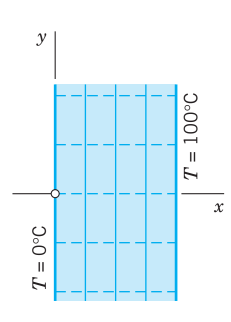


Fig. 408. Example 1

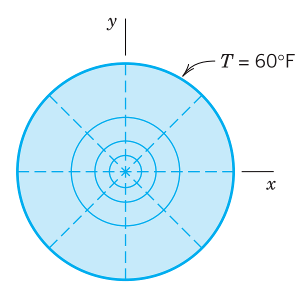


Fig. 409. Example 2

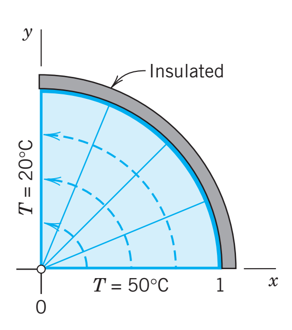


Fig. 410. Example 3

Mathematically the calculations remain the same in the transition to another field of application. Physically, new problems may arise, with boundary conditions that would make no sense physically or would be of no practical interest. This is illustrated by the next two examples.

EXAMPLE 3 A Mixed Boundary Value Problem

Find the temperature distribution in the region in Fig. 410 (cross section of a solid quarter-cylinder), whose vertical portion of the boundary is at 20°C, the horizontal portion at 50°C, and the circular portion is insulated.

Solution. The insulated portion of the boundary must be a heat flow line, since, by the insulation, heat is prevented from crossing such a curve, hence heat must flow along the curve. Thus the isotherms must meet such a curve at right angles. Since T is constant along an isotherm, this means that

(2) $$\frac{\partial T}{\partial n} = 0 \qquad \text{along an insulated portion of the boundary.}$$

Here $\partial T/\partial n$ is the **normal derivative** of T, that is, the directional derivative (Sec. 9.7) in the direction normal (perpendicular) to the insulated boundary. Such a problem in which T is prescribed on one portion of the boundary and $\partial T/\partial n$ on the other portion is called a **mixed boundary value problem**.

In our case, the normal direction to the insulated circular boundary curve is the radial direction toward the origin. Hence (2) becomes $\partial T/\partial r = 0$, meaning that along this curve the solution must not depend on r. Now $\operatorname{Arg} z = \theta$ satisfies (1), as well as this condition, and is constant (0 and $\pi/2$) on the straight portions of the boundary. Hence the solution is of the form

$$T(x, y) = a\theta + b.$$

The boundary conditions yield $a \cdot \pi/2 + b = 20$ and $a \cdot 0 + b = 50$. This gives

$$T(x, y) = 50 - \frac{60}{\pi}\,\theta, \qquad \theta = \arctan\frac{y}{x}.$$

The isotherms are portions of rays $\theta = \text{const}$. Heat flows from the x-axis along circles $r = \text{const}$ (dashed in Fig. 410) to the y-axis. ■

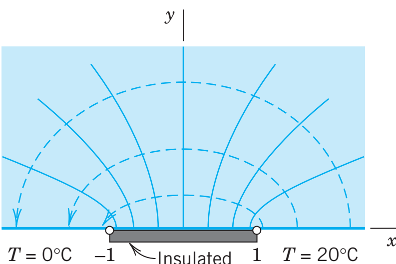


Fig. 411. Example 4: z-plane

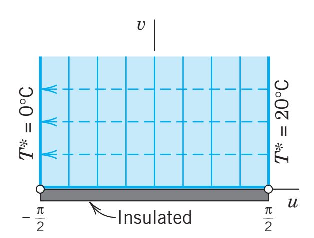


Fig. 412. Example 4: w-plane

EXAMPLE 4 Another Mixed Boundary Value Problem in Heat Conduction

Find the temperature field in the upper half-plane when the x-axis is kept at $T = 0°\text{C}$ for $x < -1$, is insulated for $-1 < x < 1$, and is kept at $T = 20°\text{C}$ for $x > 1$ (Fig. 411).

Solution. We map the half-plane in Fig. 411 onto the vertical strip in Fig. 412, find the temperature $T^*(u, v)$ there, and map it back to get the temperature $T(x, y)$ in the half-plane.

The idea of using that strip is suggested by Fig. 391 in Sec. 17.4 with the roles of $z = x + iy$ and $w = u + iv$ interchanged. The figure shows that $z = \sin w$ maps our present strip onto our half-plane in Fig. 411. Hence the inverse function

$$w = f(z) = \arcsin z$$

maps that half-plane onto the strip in the w-plane. This is the mapping function that we need according to Theorem 1 in Sec. 18.2.

The insulated segment $-1 < x < 1$ on the x-axis maps onto the segment $-\pi/2 < u < \pi/2$ on the u-axis. The rest of the x-axis maps onto the two vertical boundary portions $u = -\pi/2$ and $\pi/2$, $v > 0$, of the strip. This gives the transformed boundary conditions in Fig. 412 for $T^*(u, v)$, where on the insulated horizontal boundary, $\partial T^*/\partial n = \partial T^*/\partial v = 0$ because v is a coordinate normal to that segment.

Similarly to Example 1 we obtain

$$T^*(u, v) = 10 + \frac{20}{\pi} u$$

which satisfies all the boundary conditions. This is the real part of the complex potential $F^*(w) = 10 + (20/\pi)w$. Hence the complex potential in the z-plane is

$$F(z) = F^*(f(z)) = 10 + \frac{20}{\pi} \arcsin z$$

and $T(x, y) = \text{Re}\, F(z)$ is the solution. The isotherms are $u = \text{const}$ in the strip and the hyperbolas in the z-plane, perpendicular to which heat flows along the dashed ellipses from the 20°-portion to the cooler 0°-portion of the boundary, a physically very reasonable result. ■

Sections 18.3 and 18.5 show some of the usefulness of conformal mappings and complex potentials. Furthermore, complex potential models fluid flow in Sec. 18.4.

PROBLEM SET 18.3

1. **Parallel plates.** Find the temperature between the plates $y = 0$ and $y = d$ kept at 20 and 100°C, respectively. (i) Proceed directly. (ii) Use Example 1 and a suitable mapping.

2. **Infinite plate.** Find the temperature and the complex potential in an infinite plate with edges $y = x - 4$ and $y = x + 4$ kept at -20 and 40°C, respectively (Fig. 413). In what case will this be an approximate model?

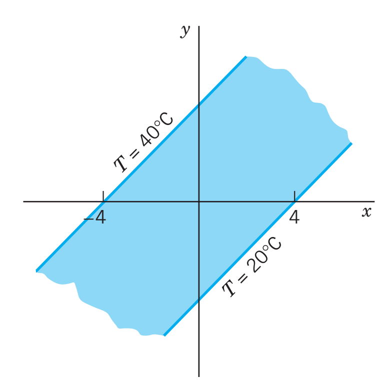


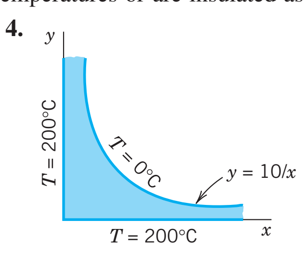
Fig. 413. Problem 2: Infinite plate

3. **CAS PROJECT. Isotherms.** Graph isotherms and lines of heat flow in Examples 2–4. Can you see from the graphs where the heat flow is very rapid?

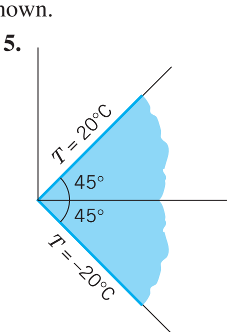
4–18 TEMPERATURE $T(x, y)$ IN PLATES

Find the temperature distribution $T(x, y)$ and the complex potential $F(z)$ in the given thin metal plate whose faces are insulated and whose edges are kept at the indicated temperatures or are insulated as shown.

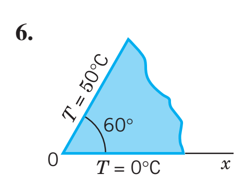
4.



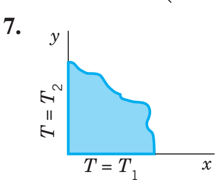
5.



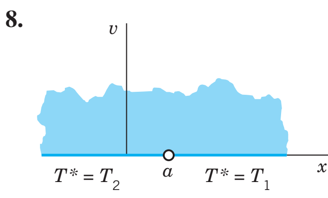
6.



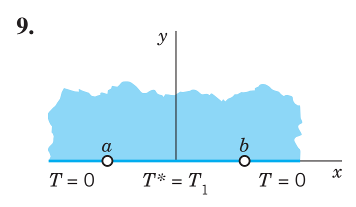
7.



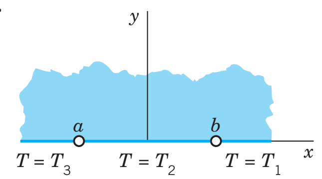
8.



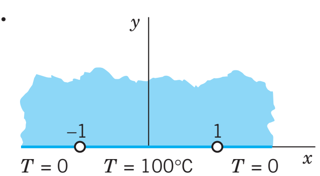
9.



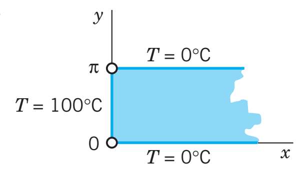
10.



11.

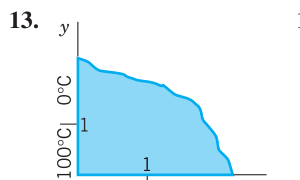


12.

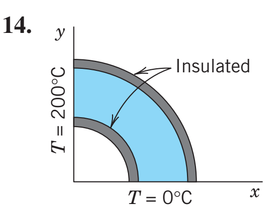


Hint. Apply $w = \cosh z$ to Prob. 11.

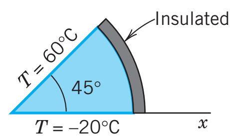
13.



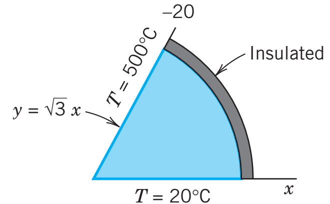
14.



15.



16.



17. First quadrant of the z-plane with y-axis kept at 100°C, the segment $0 < x < 1$ of the x-axis insulated and the x-axis for $x > 1$ kept at 200°C. *Hint.* Use Example 4.

18. Figure 410, $T(0, y) = -30°\text{C}$, $T(x, 0) = 100°\text{C}$

19. **Interpretation.** Formulate Prob. 11 in terms of electrostatics.

20. **Interpretation.** Interpret Prob. 17 in Sec. 18.2 as a heat problem, with boundary temperatures, say, 10°C on the upper part and 200°C on the lower.

18.4 Fluid Flow

Laplace's equation also plays a basic role in hydrodynamics, in steady nonviscous fluid flow under physical conditions discussed later in this section. For methods of complex analysis to be applicable, our problems will be *two-dimensional*, so that the **velocity vector** V by which the motion of the fluid can be given depends only on two space variables x and y, and the motion is the same in all planes parallel to the xy-plane.

Then we can use for the velocity vector V a complex function

$$V = V_1 + iV_2 \tag{1}$$

giving the magnitude $|V|$ and direction Arg V of the velocity at each point $z = x + iy$. Here V_1 and V_2 are the components of the velocity in the x and y directions. V is tangential to the path of the moving particles, called a **streamline** of the motion (Fig. 414).

We show that under suitable assumptions (explained in detail following the examples), for a given flow there exists an analytic function

$$F(z) = \Phi(x, y) + i\Psi(x, y), \tag{2}$$

called the **complex potential** of the flow, such that the streamlines are given by $\Psi(x, y) = \text{const}$, and the velocity vector or, briefly, the **velocity** is given by

$$V = V_1 + iV_2 = \overline{F'(z)} \tag{3}$$

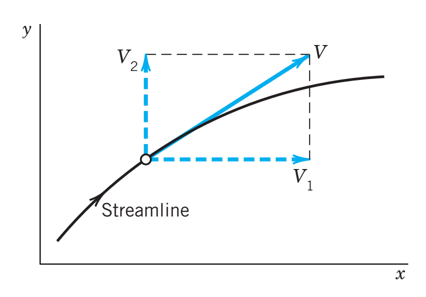


Fig. 414. Velocity

where the bar denotes the complex conjugate. Ψ is called the **stream function**. The function Φ is called the **velocity potential**. The curves $\Phi(x, y) = \text{const}$ are called **equipotential lines**. The velocity vector V is the **gradient** of Φ; by definition, this means that

$$V_1 = \frac{\partial \Phi}{\partial x}, \qquad V_2 = \frac{\partial \Phi}{\partial y}. \tag{4}$$

Indeed, for $F = \Phi + i\Psi$, Eq. (4) in Sec. 13.4 is $F' = \Phi_x + i\Psi_x$ with $\Psi_x = -\Phi_y$ by the second Cauchy–Riemann equation. Together we obtain (3):

$$\overline{F'(z)} = \Phi_x - i\Psi_x = \Phi_x + i\Phi_y = V_1 + iV_2 = V.$$

Furthermore, since $F(z)$ is analytic, Φ and Ψ satisfy Laplace's equation

$$(5) \qquad \nabla^2\Phi = \frac{\partial^2\Phi}{\partial x^2} + \frac{\partial^2\Phi}{\partial y^2} = 0, \qquad \nabla^2\Psi = \frac{\partial^2\Psi}{\partial x^2} + \frac{\partial^2\Psi}{\partial y^2} = 0.$$

Whereas in electrostatics the boundaries (conducting plates) are equipotential lines, in fluid flow the boundaries across which fluid cannot flow must be streamlines. Hence in fluid flow the stream function is of particular importance.

Before discussing the conditions for the validity of the statements involving (2)–(5), let us consider two flows of practical interest, so that we first see what is going on from a practical point of view. Further flows are included in the problem set.

EXAMPLE 1 Flow Around a Corner

The complex potential $F(z) = z^2 = x^2 - y^2 + 2ixy$ models a flow with

Equipotential lines	$\Phi = x^2 - y^2 = \text{const}$	(Hyperbolas)
Streamlines	$\Psi = 2xy = \text{const}$	(Hyperbolas).

From (3) we obtain the velocity vector

$$V = 2\bar{z} = 2(x - iy), \qquad \text{that is,} \qquad V_1 = 2x, \qquad V_2 = -2y.$$

The speed (magnitude of the velocity) is

$$|V| = \sqrt{V_1^2 + V_2^2} = 2\sqrt{x^2 + y^2}.$$

The flow may be interpreted as the flow in a channel bounded by the positive coordinates axes and a hyperbola, say, $xy = 1$ (Fig. 415). We note that the speed along a streamline S has a minimum at the point P where the cross section of the channel is large. ■

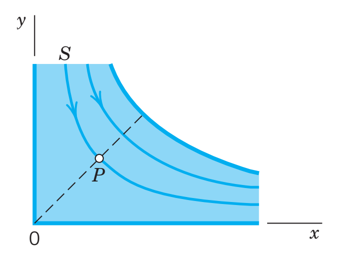


Fig. 415. Flow around a corner (Example 1)

EXAMPLE 2 Flow Around a Cylinder

Consider the complex potential

$$F(z) = \Phi(x, y) + i\Psi(x, y) = z + \frac{1}{z}.$$

Using the polar form $z = re^{i\theta}$, we obtain

$$F(z) = re^{i\theta} + \frac{1}{r}e^{-i\theta} = \left(r + \frac{1}{r}\right)\cos\theta + i\left(r - \frac{1}{r}\right)\sin\theta.$$

Hence the streamlines are

$$\Psi(x, y) = \left(r - \frac{1}{r}\right)\sin\theta = \text{const}.$$

In particular, $\Psi(x, y) = 0$ gives $r - 1/r = 0$ or $\sin\theta = 0$. Hence this streamline consists of the unit circle ($r = 1/r$ gives $r = 1$) and the x-axis ($\theta = 0$ and $\theta = \pi$). For large $|z|$ the term $1/z$ in $F(z)$ is small in absolute value, so that for these z the flow is nearly uniform and parallel to the x-axis. Hence we can interpret this as a flow around a long circular cylinder of unit radius that is perpendicular to the z-plane and intersects it in the unit circle $|z| = 1$ and whose axis corresponds to $z = 0$.

The flow has two **stagnation points** (that is, points at which the velocity V is zero), at $z = \pm 1$. This follows from (3) and

$$F'(z) = 1 - \frac{1}{z^2}, \qquad \text{hence} \qquad z^2 - 1 = 0. \qquad \text{(See Fig. 416.)}$$

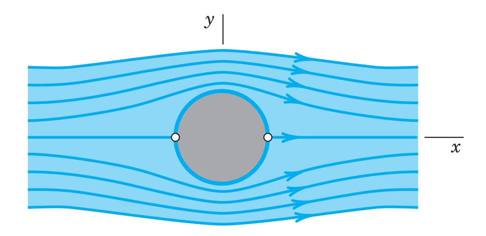


Fig. 416. Flow around a cylinder (Example 2)

Assumptions and Theory Underlying (2)–(5)

THEOREM 1

Complex Potential of a Flow

If the domain of flow is simply connected and the flow is irrotational and incompressible, then the statements involving (2)–(5) *hold. In particular, then the flow has a complex potential* $F(z)$*, which is an analytic function.* (Explanation of terms below.)

PROOF We prove this theorem, along with a discussion of basic concepts related to fluid flow.

(a) ***First Assumption: Irrotational.*** Let C be any smooth curve in the z-plane given by $z(s) = x(s) + iy(s)$, where s is the arc length of C. Let the real variable V_t be the component of the velocity V tangent to C (Fig. 417). Then the value of the real line integral

$$\int_C V_t \, ds \tag{6}$$

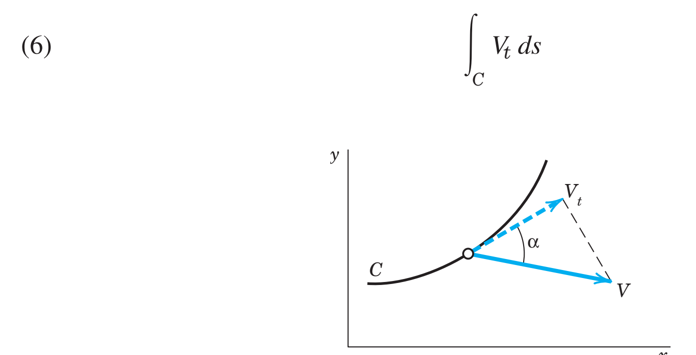


Fig. 417. Tangential component of the velocity with respect to a curve C

taken along C in the sense of increasing s is called the **circulation** of the fluid along C, a name that will be motivated as we proceed in this proof. Dividing the circulation by the length of C, we obtain the *mean velocity*[1] of the flow along the curve C. Now

$$V_t = |V| \cos \alpha \qquad \text{(Fig. 417).}$$

Hence V_t is the dot product (Sec. 9.2) of V and the tangent vector dz/ds of C (Sec. 17.1); thus in (6),

$$V_t \, ds = \left(V_1 \frac{dx}{ds} + V_2 \frac{dy}{ds} \right) ds = V_1 \, dx + V_2 \, dy.$$

The circulation (6) along C now becomes

$$\int_C V_t \, ds = \int_C (V_1 \, dx + V_2 \, dy). \tag{7}$$

As the next idea, let C be a ***closed curve*** satisfying the assumption as in Green's theorem (Sec. 10.4), and let C be the boundary of a simply connected domain D. Suppose further that V has continuous partial derivatives in a domain containing D and C. Then we can use Green's theorem to represent the circulation around C by a double integral,

$$\oint_C (V_1 \, dx + V_2 \, dy) = \iint_D \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \right) dx \, dy. \tag{8}$$

The integrand of this double integral is called the **vorticity** of the flow. The vorticity divided by 2 is called the **rotation**

$$\omega(x, y) = \frac{1}{2} \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} \right). \tag{9}$$

We assume the flow to be **irrotational**, that is, $\omega(x, y) \equiv 0$ throughout the flow; thus,

$$\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y} = 0. \tag{10}$$

To understand the physical meaning of vorticity and rotation, take for C in (8) a circle. Let r be the radius of C. Then the circulation divided by the length $2\pi r$ of C is the mean

---

[1] *Definitions:* $\dfrac{1}{b-a} \displaystyle\int_a^b f(x) \, dx =$ **mean value** of f on the interval $a \leqq x \leqq b$,

$\dfrac{1}{L} \displaystyle\int_C f(s) \, ds =$ **mean value** of f on C $\qquad$ (L = length of C),

$\dfrac{1}{A} \displaystyle\iint_D f(x, y) \, dx \, dy =$ **mean value** of f on D $\qquad$ (A = area of D).

velocity of the fluid along C. Hence by dividing this by r we obtain the mean **angular velocity** ω_0 of the fluid about the center of the circle:

$$\omega_0 = \frac{1}{2\pi r^2} \iint_D \left(\frac{\partial V_2}{\partial x} - \frac{\partial V_1}{\partial y}\right) dx\, dy = \frac{1}{\pi r^2} \iint_D \omega(x, y)\, dx\, dy.$$

If we now let $r \to 0$, the limit of ω_0 is the value of ω at the center of C. Hence $\omega(x, y)$ is the limiting angular velocity of a circular element of the fluid as the circle shrinks to the point (x, y). Roughly speaking, *if a spherical element of the fluid were suddenly solidified and the surrounding fluid simultaneously annihilated, the element would rotate with the angular velocity* ω.

(b) ***Second Assumption: Incompressible.*** Our second assumption is that the fluid is incompressible. (Fluids include liquids, which are incompressible, and gases, such as air, which are compressible.) Then

$$\frac{\partial V_1}{\partial x} + \frac{\partial V_2}{\partial y} = 0 \tag{11}$$

in every region that is free of **sources** or **sinks**, that is, points at which fluid is produced or disappears, respectively. The expression in (11) is called the **divergence** of V and is denoted by div V. (See also (7) in Sec. 9.8.)

(c) ***Complex Velocity Potential.*** If the domain D of the flow is simply connected (Sec. 14.2) and the flow is irrotational, then (10) implies that the line integral (7) is independent of path in D (by Theorem 3 in Sec. 10.2, where $F_1 = V_1$, $F_2 = V_2$, $F_3 = 0$, and z is the third coordinate in space and has nothing to do with our present z). Hence if we integrate from a fixed point (a, b) in D to a variable point (x, y) in D, the integral becomes a function of the point (x, y), say, $\Phi(x, y)$:

$$\Phi(x, y) = \int_{(a,b)}^{(x,y)} (V_1\, dx + V_2\, dy). \tag{12}$$

We claim that the flow has a velocity potential Φ, which is given by (12). To prove this, all we have to do is to show that (4) holds. Now since the integral (7) is independent of path, $V_1\, dx + V_2\, dy$ is exact (Sec. 10.2), namely, the differential of Φ, that is,

$$V_1\, dx + V_2\, dy = \frac{\partial \Phi}{\partial x}\, dx + \frac{\partial \Phi}{\partial y}\, dy.$$

From this we see that $V_1 = \partial\Phi/\partial x$ and $V_2 = \partial\Phi/\partial y$, which gives (4).

That Φ is harmonic follows at once by substituting (4) into (11), which gives the first Laplace equation in (5).

We finally take a harmonic conjugate Ψ of Φ. Then the other equation in (5) holds. Also, since the second partial derivatives of Φ and Ψ are continuous, we see that the complex function

$$F(z) = \Phi(x, y) + i\Psi(x, y)$$

is analytic in D. Since the curves $\Psi(x, y) = \text{const}$ are perpendicular to the equipotential curves $\Phi(x, y) = \text{const}$ (except where $F'(z) = 0$), we conclude that $\Psi(x, y) = \text{const}$ are the streamlines. Hence Ψ is the stream function and $F(z)$ is the complex potential of the flow. This completes the proof of Theorem 1 as well as our discussion of the important role of complex analysis in compressible fluid flow. ■

PROBLEM SET 18.4

1. **Differentiability.** Under what condition on the velocity vector V in (1) will $F(z)$ in (2) be analytic?
2. **Corner flow.** Along what curves will the speed in Example 1 be constant? Is this obvious from Fig. 415?
3. **Cylinder.** Guess from physics and from Fig. 416 where on the y-axis the speed is maximum. Then calculate.
4. **Cylinder.** Calculate the speed along the cylinder wall in Fig. 416, also confirming the answer to Prob. 3.
5. **Irrotational flow.** Show that the flow in Example 2 is irrotational.
6. **Extension of Example 1.** Sketch or graph and interpret the flow in Example 1 on the whole upper half-plane.
7. **Parallel flow.** Sketch and interpret the flow with complex potential $F(z) = z$.
8. **Parallel flow.** What is the complex potential of an upward parallel flow of speed $K > 0$ in the direction of $y = x$? Sketch the flow.
9. **Corner.** What $F(z)$ would be suitable in Example 1 if the angle of the corner were $\pi/4$ instead of $\pi/2$?
10. **Corner.** Show that $F(z) = iz^2$ also models a flow around a corner. Sketch streamlines and equipotential lines. Find V.
11. What flow do you obtain from $F(z) = -iKz$, K positive real?
12. **Conformal mapping.** Obtain the flow in Example 1 from that in Prob. 11 by a suitable conformal mapping.
13. **60°- Sector.** What $F(z)$ would be suitable in Example 1 if the angle at the corner were $\pi/3$?
14. Sketch or graph streamlines and equipotential lines of $F(z) = iz^3$. Find V. Find all points at which V is horizontal.
15. Change $F(z)$ in Example 2 slightly to obtain a flow around a cylinder of radius r_0 that gives the flow in Example 2 if $r_0 \to 1$.
16. **Cylinder.** What happens in Example 2 if you replace z by z^2? Sketch and interpret the resulting flow in the first quadrant.
17. **Elliptic cylinder.** Show that $F(z) = \arccos z$ gives confocal ellipses as streamlines, with foci at $z = \pm 1$, and that the flow circulates around an elliptic cylinder or a plate (the segment from -1 to 1 in Fig. 418).

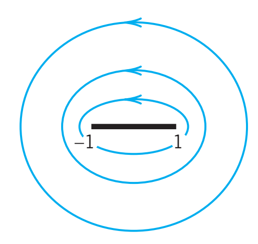


Fig. 418. Flow around a plate in Prob. 17.

18. **Aperture.** Show that $F(z) = \operatorname{arccosh} z$ gives confocal hyperbolas as streamlines, with foci at $z = \pm 1$, and the flow may be interpreted as a flow through an aperture (Fig. 419).

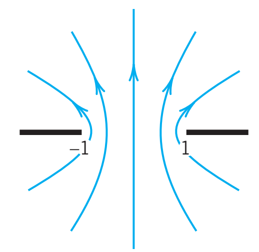


Fig. 419. Flow through an aperture in Prob. 18.

19. **Potential $F(z) = 1/z$.** Show that the streamlines of $F(z) = 1/z$ and circles through the origin with centers on the y-axis.
20. **TEAM PROJECT. Role of the Natural Logarithm in Modeling Flows. (a) Basic flows: Source and sink.** Show that $F(z) = (c/2\pi) \ln z$ with constant positive real c gives a flow directed radially outward (Fig. 420), so that F models a **point source** at $z = 0$ (that is, a **source line** $x = 0, y = 0$ in space) at which fluid is produced. c is called the **strength** or **discharge** of the source. If c is negative real, show that the flow is directed radially inward, so that F models a **sink** at $z = 0$, a point at which fluid disappears. Note that $z = 0$ is the singular point of $F(z)$.

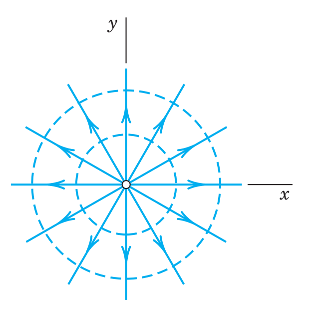


Fig. 420. Point source

(b) Basic flows: Vortex. Show that $F(z) = -(Ki/2\pi) \ln z$ with positive real K gives a flow circulating counterclockwise around $z = 0$ (Fig. 421). $z = 0$ is called a **vortex**. Note that each time we travel around the vortex, the potential increases by K.

(c) Addition of flows. Show that addition of the velocity vectors of two flows gives a flow whose complex potential is obtained by adding the complex potentials of those flows.

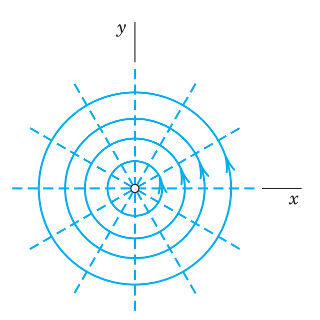


Fig. 421. Vortex flow

(d) Source and sink combined. Find the complex potentials of a flow with a source of strength 1 at $z = -a$ and of a flow with a sink of strength 1 at $z = a$. Add both and sketch or graph the streamlines. Show that for small $|a|$ these lines look similar to those in Prob. 19.

(e) Flow with circulation around a cylinder. Add the potential in (b) to that in Example 2. Show that this gives a flow for which the cylinder wall $|z| = 1$ is a streamline. Find the speed and show that the stagnation points are

$$z = \frac{iK}{4\pi} \pm \sqrt{\frac{-K^2}{16\pi^2} + 1};$$

if $K = 0$ they are at ± 1; as K increases they move up on the unit circle until they unite at $z = i$ ($K = 4\pi$, see Fig. 422), and if $K > 4\pi$ they lie on the imaginary axis (one lies in the field of flow and the other one lies inside the cylinder and has no physical meaning).

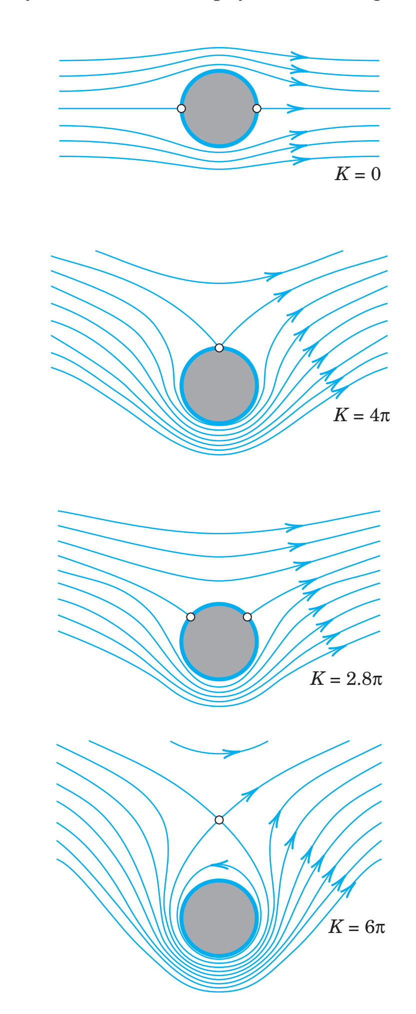


Fig. 422. Flow around a cylinder without circulation ($K = 0$) and with circulation

18.5 Poisson's Integral Formula for Potentials

So far in this chapter we have seen powerful methods based on conformal mappings and complex potentials. They were used for modeling and solving two-dimensional potential problems and demonstrated the importance of complex analysis.

Now we introduce a further method that results from complex integration. It will yield the very important Poisson integral formula (5) for potentials in a standard domain

(a circular disk). In addition, from (5), we will derive a useful series (7) for these potentials. This allows us to solve problems for disks and then map solutions conformally onto other domains.

Derivation of Poisson's Integral Formula

Poisson's formula will follow from Cauchy's integral formula (Sec. 14.3)

$$F(z) = \frac{1}{2\pi i} \oint_C \frac{F(z^*)}{z^* - z} \, dz^*. \tag{1}$$

Here C is the circle $z^* = Re^{i\alpha}$ (counterclockwise, $0 \leqq \alpha \leqq 2\pi$), and we assume that $F(z^*)$ is analytic in a domain containing C and its full interior. Since $dz^* = iRe^{i\alpha}\, d\alpha = iz^*\, d\alpha$, we obtain from (1)

$$F(z) = \frac{1}{2\pi} \int_0^{2\pi} F(z^*) \frac{z^*}{z^* - z} \, d\alpha \qquad (z^* = Re^{i\alpha},\ z = re^{i\theta}). \tag{2}$$

Now comes a little trick. If instead of z inside C we take a Z outside C, the integrals (1) and (2) are zero by Cauchy's integral theorem (Sec. 14.2). We choose $Z = z^*\bar{z}^*/\bar{z} = R^2/\bar{z}$, which is outside C because $|Z| = R^2/|z| = R^2/r > R$. From (2) we thus have

$$0 = \frac{1}{2\pi} \int_0^{2\pi} F(z^*) \frac{z^*}{z^* - Z} \, d\alpha = \frac{1}{2\pi} \int_0^{2\pi} F(z^*) \frac{z^*}{z^* - \dfrac{z^*\bar{z}^*}{\bar{z}}} \, d\alpha$$

and by straightforward simplification of the last expression on the right,

$$0 = \frac{1}{2\pi} \int_0^{2\pi} F(z^*) \frac{\bar{z}}{\bar{z} - \bar{z}^*} \, d\alpha.$$

We subtract this from (2) and use the following formula that you can verify by direct calculation ($\bar{z}z^*$ cancels):

$$\frac{z^*}{z^* - z} - \frac{\bar{z}}{\bar{z} - \bar{z}^*} = \frac{z^*\bar{z}^* - z\bar{z}}{(z^* - z)(\bar{z}^* - \bar{z})}. \tag{3}$$

We then have

$$F(z) = \frac{1}{2\pi} \int_0^{2\pi} F(z^*) \frac{z^*\bar{z}^* - z\bar{z}}{(z^* - z)(\bar{z}^* - \bar{z})} \, d\alpha. \tag{4}$$

From the polar representations of z and z^* we see that the quotient in the integrand is real and equal to

$$\frac{R^2 - r^2}{(Re^{i\alpha} - re^{i\theta})(Re^{-i\alpha} - re^{-i\theta})} = \frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \alpha) + r^2}.$$

We now write $F(z) = \Phi(r, \theta) + i\Psi(r, \theta)$ and take the real part on both sides of (4). Then we obtain **Poisson's integral formula**[2]

(5)
$$\Phi(r, \theta) = \frac{1}{2\pi}\int_0^{2\pi} \Phi(R, \alpha)\,\frac{R^2 - r^2}{R^2 - 2Rr\cos(\theta - \alpha) + r^2}\,d\alpha.$$

This formula represents the harmonic function Φ in the disk $|z| \leqq R$ in terms of its values $\Phi(R, \alpha)$ on the boundary (the circle) $|z| = R$.

Formula (5) is still valid if the boundary function $\Phi(R, \alpha)$ is merely piecewise continuous (as is practically often the case; see Figs. 405 and 406 in Sec. 18.2 for an example). Then (5) gives a function harmonic in the open disk, and on the circle $|z| = R$ equal to the given boundary function, except at points where the latter is discontinuous. A proof can be found in Ref. [D1] in App. 1.

Series for Potentials in Disks

From (5) we may obtain an important series development of Φ in terms of simple harmonic functions. We remember that the quotient in the integrand of (5) was derived from (3). We claim that the right side of (3) is the real part of

$$\frac{z^* + z}{z^* - z} = \frac{(z^* + z)(\bar{z}^* - \bar{z})}{(z^* - z)(\bar{z}^* - \bar{z})} = \frac{z^*\bar{z}^* - z\bar{z} - z^*\bar{z} + z\bar{z}^*}{|z^* - z|^2}.$$

Indeed, the last denominator is real and so is $z^*\bar{z}^* - z\bar{z}$ in the numerator, whereas $-z^*\bar{z} + z\bar{z}^* = 2i\,\mathrm{Im}\,(z\bar{z}^*)$ in the numerator is pure imaginary. This verifies our claim. Now by the use of the geometric series we obtain (develop the denominator)

(6)
$$\frac{z^* + z}{z^* - z} = \frac{1 + (z/z^*)}{1 - (z/z^*)} = \left(1 + \frac{z}{z^*}\right)\sum_{n=0}^{\infty}\left(\frac{z}{z^*}\right)^n = 1 + 2\sum_{n=1}^{\infty}\left(\frac{z}{z^*}\right)^n.$$

Since $z = re^{i\theta}$ and $z^* = Re^{i\alpha}$, we have

$$\mathrm{Re}\left[\left(\frac{z}{z^*}\right)^n\right] = \mathrm{Re}\left[\frac{r^n}{R^n}\,e^{in\theta}e^{-in\alpha}\right] = \left(\frac{r}{R}\right)^n \cos(n\theta - n\alpha).$$

On the right, $\cos(n\theta - n\alpha) = \cos n\theta \cos n\alpha + \sin n\theta \sin n\alpha$. Hence from (6) we obtain

(6*)
$$\begin{aligned}\mathrm{Re}\,\frac{z^* + z}{z^* - z} &= 1 + 2\sum_{n=1}^{\infty}\mathrm{Re}\left(\frac{z}{z^*}\right)^n \\ &= 1 + 2\sum_{n=1}^{\infty}\left(\frac{r}{R}\right)^n(\cos n\theta\cos n\alpha + \sin n\theta\sin n\alpha).\end{aligned}$$

[2] SIMÉON DENIS POISSON (1781–1840), French mathematician and physicist, professor in Paris from 1809. His work includes potential theory, partial differential equations (Poisson equation, Sec. 12.1), and probability (Sec. 24.7).

This expression is equal to the quotient in (5), as we have mentioned before, and by inserting it into (5) and integrating term by term with respect to α from 0 to 2π we obtain

$$\Phi(r, \theta) = a_0 + \sum_{n=1}^{\infty} \left(\frac{r}{R}\right)^n (a_n \cos n\theta + b_n \sin n\theta) \tag{7}$$

where the coefficients are [the 2 in (6*) cancels the 2 in $1/(2\pi)$ in (5)]

$$a_0 = \frac{1}{2\pi}\int_0^{2\pi} \Phi(R, \alpha)\, d\alpha, \qquad a_n = \frac{1}{\pi}\int_0^{2\pi} \Phi(R, \alpha) \cos n\alpha\, d\alpha,$$

$$(8) \qquad\qquad n = 1, 2, \cdots,$$

$$b_n = \frac{1}{\pi}\int_0^{2\pi} \Phi(R, \alpha) \sin n\alpha\, d\alpha,$$

the Fourier coefficients of $\Phi(R, \alpha)$; see Sec. 11.1. Now, for $r = R$, the series (7) becomes the Fourier series of $\Phi(R, \alpha)$. Hence the representation (7) will be valid whenever the given $\Phi(R, \alpha)$ on the boundary can be represented by a Fourier series.

EXAMPLE 1 Dirichlet Problem for the Unit Disk

Find the electrostatic potential $\Phi(r, \theta)$ in the unit disk $r < 1$ having the boundary values

$$\Phi(1, \alpha) = \begin{cases} -\alpha/\pi & \text{if } -\pi < \alpha < 0 \\ \alpha/\pi & \text{if } 0 < \alpha < \pi \end{cases} \qquad \text{(Fig. 423).}$$

Solution. Since $\Phi(1, \alpha)$ is even, $b_n = 0$, and from (8) we obtain $a_0 = \frac{1}{2}$ and

$$a_n = \frac{1}{\pi}\left[-\int_{-\pi}^{0} \frac{\alpha}{\pi} \cos n\alpha\, d\alpha + \int_0^{\pi} \frac{\alpha}{\pi} \cos n\alpha\, d\alpha\right] = \frac{2}{n^2\pi^2}(\cos n\pi - 1).$$

Hence, $a_n = -4/(n^2\pi^2)$ if n is odd, $a_n = 0$ if $n = 2, 4, \cdots$, and the potential is

$$\Phi(r, \theta) = \frac{1}{2} - \frac{4}{\pi^2}\left[r \cos\theta + \frac{r^3}{3^2} \cos 3\theta + \frac{r^5}{5^2} \cos 5\theta + \cdots\right].$$

Figure 424 shows the unit disk and some of the equipotential lines (curves Φ = const). ■

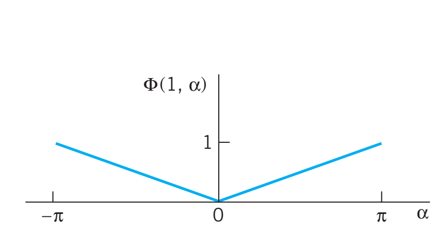


Fig. 423. Boundary values in Example 1

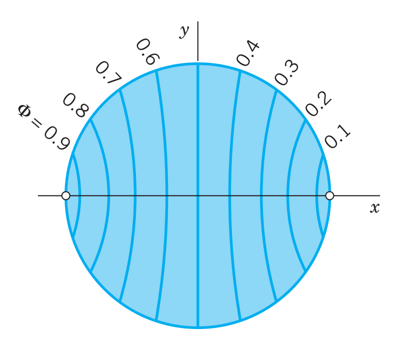


Fig. 424. Potential in Example 1

PROBLEM SET 18.5

1. Give the details of the derivation of the series (7) from the Poisson formula (5).
2. Verify (3).
3. Show that each term of (7) is a harmonic function in the disk $r < R$.
4. Why does the series in Example 1 reduce to a **cosine series?**

5–18 HARMONIC FUNCTIONS IN A DISK

Using (7), find the potential $\Phi(r, \theta)$ in the unit disk $r < 1$ having the given boundary values $\Phi(1, \theta)$. Using the sum of the first few terms of the series, compute some values of Φ and sketch a figure of the equipotential lines.

5. $\Phi(1, \theta) = \frac{3}{2} \sin 3\theta$
6. $\Phi(1, \theta) = 5 - \cos 2\theta$
7. $\Phi(1, \theta) = a \cos^2 4\theta$
8. $\Phi(1, \theta) = 4 \sin^3 \theta$
9. $\Phi(1, \theta) = 8 \sin^4 \theta$
10. $\Phi(1, \theta) = 16 \cos^3 2\theta$
11. $\Phi(1, \theta) = \theta/\pi \quad \text{if } -\pi < \theta < \pi$
12. $\Phi(1, \theta) = k \quad \text{if } 0 < \theta < \pi$ and 0 otherwise
13. $\Phi(1, \theta) = \theta \quad \text{if } -\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise
14. $\Phi(1, \theta) = |\theta|/\pi \quad \text{if } -\pi < \theta < \pi$
15. $\Phi(1, \theta) = 1 \quad \text{if } -\frac{1}{2}\pi < \theta < \frac{1}{2}\pi$ and 0 otherwise
16. $\Phi(1, \theta) = \begin{cases} \theta + \pi & \text{if } -\pi < \theta < 0 \\ \theta - \pi & \text{if } 0 < \theta < \pi \end{cases}$
17. $\Phi(1, \theta) = \theta^2/\pi^2 \quad \text{if } -\pi < \theta < \pi$
18. $\Phi(1, \theta) = \begin{cases} 0 & \text{if } -\pi < \theta < 0 \\ \theta & \text{if } 0 < \theta < \pi \end{cases}$
19. **CAS EXPERIMENT. Series (7).** Write a program for series developments (7). Experiment on accuracy by computing values from partial sums and comparing them with values that you obtain from your CAS graph. Do this **(a)** for Example 1 and Fig. 424, **(b)** for Φ in Prob. 11 (which is discontinuous on the boundary!), **(c)** for a Φ of your choice with continuous boundary values, and **(d)** for Φ with discontinuous boundary values.
20. **TEAM PROJECT. Potential in a Disk. (a) Mean value property.** Show that the value of a harmonic function Φ at the center of a circle C equals the mean of the value of Φ on C (see Sec. 18.4, footnote 1, for definitions of mean values).

    (b) Separation of variables. Show that the terms of (7) appear as solutions in separating the Laplace equation in polar coordinates.

    (c) Harmonic conjugate. Find a series for a harmonic conjugate Ψ of Φ from (7). *Hint.* Use the Cauchy–Riemann equations.

    (d) Power series. Find a series for $F(z) = \Phi + i\Psi$.

18.6 General Properties of Harmonic Functions. Uniqueness Theorem for the Dirichlet Problem

Recall from Sec. 10.8 that harmonic functions are solutions to Laplace's equation and their second-order partial derivatives are continuous. In this section we explore how general properties of harmonic functions often can be obtained from properties of analytic functions. This can frequently be done in a simple fashion. Specifically, important mean value properties of harmonic functions follow readily from those of analytic functions. The details are as follows.

THEOREM 1

Mean Value Property of Analytic Functions

Let $F(z)$ be analytic in a simply connected domain D. Then the value of $F(z)$ at a point z_0 in D is equal to the mean value of $F(z)$ on any circle in D with center at z_0.

PROOF In Cauchy's integral formula (Sec. 14.3)

$$F(z_0) = \frac{1}{2\pi i} \oint_C \frac{F(z)}{z - z_0}\, dz \tag{1}$$

we choose for C the circle $z = z_0 + re^{i\alpha}$ in D. Then $z - z_0 = re^{i\alpha}$, $dz = ire^{i\alpha}\, d\alpha$, and (1) becomes

$$F(z_0) = \frac{1}{2\pi} \int_0^{2\pi} F(z_0 + re^{i\alpha})\, d\alpha. \tag{2}$$

The right side is the mean value of F on the circle (= value of the integral divided by the length 2π of the interval of integration). This proves the theorem. ∎

For harmonic functions, Theorem 1 implies

THEOREM 2

Two Mean Value Properties of Harmonic Functions

Let $\Phi(x, y)$ be harmonic in a simply connected domain D. Then the value of $\Phi(x, y)$ at a point (x_0, y_0) in D is equal to the mean value of $\Phi(x, y)$ on any circle in D with center at (x_0, y_0). This value is also equal to the mean value of $\Phi(x, y)$ on any circular disk in D with center (x_0, y_0). [See footnote 1 in Sec. 18.4.]

PROOF The first part of the theorem follows from (2) by taking the real parts on both sides,

$$\Phi(x_0, y_0) = \operatorname{Re} F(x_0 + iy_0) = \frac{1}{2\pi} \int_0^{2\pi} \Phi(x_0 + r\cos\alpha, y_0 + r\sin\alpha)\, d\alpha.$$

The second part of the theorem follows by integrating this formula over r from 0 to r_0 (the radius of the disk) and dividing by $r_0^2/2$,

$$\Phi(x_0, y_0) = \frac{1}{\pi r_0^2} \int_0^{r_0} \int_0^{2\pi} \Phi(x_0 + r\cos\alpha, y_0 + r\sin\alpha) r\, d\alpha\, dr. \tag{3}$$

The right side is the indicated mean value (integral divided by the area of the region of integration). ∎

Returning to analytic functions, we state and prove another famous consequence of Cauchy's integral formula. The proof is indirect and shows quite a nice idea of applying the *ML*-inequality. (A *bounded region* is a region that lies entirely in some circle about the origin.)

THEOREM 3

Maximum Modulus Theorem for Analytic Functions

Let $F(z)$ be analytic and nonconstant in a domain containing a bounded region R and its boundary. Then the absolute value $|F(z)|$ cannot have a maximum at an interior point of R. Consequently, the maximum of $|F(z)|$ is taken on the boundary of R. If $F(z) \neq 0$ in R, the same is true with respect to the minimum of $|F(z)|$.

PROOF We assume that $|F(z)|$ has a maximum at an interior point z_0 of R and show that this leads to a contradiction. Let $|F(z_0)| = M$ be this maximum. Since $F(z)$ is not constant, $|F(z)|$ is not constant, as follows from Example 3 in Sec. 13.4. Consequently, we can find a circle C of radius r with center at z_0 such that the interior of C is in R and $|F(z)|$ is smaller than M at some point P of C. Since $|F(z)|$ is continuous, it will be smaller than M on an arc C_1 of C that contains P (see Fig. 425), say,

$$|F(z)| \leqq M - k \quad (k > 0) \qquad \text{for all } z \text{ on } C_1.$$

Let C_1 have the length L_1. Then the complementary arc C_2 of C has the length $2\pi r - L_1$. We now apply the ML-inequality (Sec. 14.1) to (1) and note that $|z - z_0| = r$. We then obtain (using straightforward calculation in the second line of the formula)

$$\begin{aligned} M = |F(z_0)| &\leqq \frac{1}{2\pi}\left|\int_{C_1} \frac{F(z)}{z - z_0}\,dz\right| + \frac{1}{2\pi}\left|\int_{C_2} \frac{F(z)}{z - z_0}\,dz\right| \\ &\leqq \frac{1}{2\pi}\left(\frac{M-k}{r}\right)L_1 + \frac{1}{2\pi}\left(\frac{M}{r}\right)(2\pi r - L_1) = M - \frac{kL_1}{2\pi r} < M \end{aligned}$$

that is, $M < M$, which is impossible. Hence our assumption is false and the first statement is proved.

Next we prove the second statement. If $F(z) \neq 0$ in R, then $1/F(z)$ is analytic in R. From the statement already proved it follows that the maximum of $1/|F(z)|$ lies on the boundary of R. But this maximum corresponds to the minimum of $|F(z)|$. This completes the proof. ■

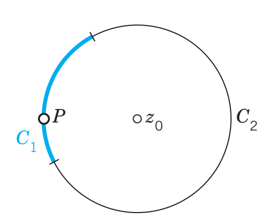


Fig. 425. Proof of Theorem 3

This theorem has several fundamental consequences for harmonic functions, as follows.

THEOREM 4

Harmonic Functions

Let $\Phi(x, y)$ be harmonic in a domain containing a simply connected bounded region R and its boundary curve C. Then:

(I) **(Maximum principle)** *If $\Phi(x, y)$ is not constant, it has neither a maximum nor a minimum in R. Consequently, the maximum and the minimum are taken on the boundary of R.*

(II) *If $\Phi(x, y)$ is constant on C, then $\Phi(x, y)$ is a constant.*

(III) *If $h(x, y)$ is harmonic in R and on C and if $h(x, y) = \Phi(x, y)$ on C, then $h(x, y) = \Phi(x, y)$ everywhere in R.*

PROOF **(I)** Let $\Psi(x, y)$ be a conjugate harmonic function of $\Phi(x, y)$ in R. Then the complex function $F(z) = \Phi(x, y) + i\Psi(x, y)$ is analytic in R, and so is $G(z) = e^{F(z)}$. Its absolute value is

$$|G(z)| = e^{\text{Re}\, F(z)} = e^{\Phi(x, y)}.$$

From Theorem 3 it follows that $|G(z)|$ cannot have a maximum at an interior point of R. Since e^{Φ} is a monotone increasing function of the real variable Φ, the statement about the maximum of Φ follows. From this, the statement about the minimum follows by replacing Φ by $-\Phi$.

(II) By (I) the function $\Phi(x, y)$ takes its maximum and its minimum on C. Thus, if $\Phi(x, y)$ is constant on C, its minimum must equal its maximum, so that $\Phi(x, y)$ must be a constant.

(III) If h and Φ are harmonic in R and on C, then $h - \Phi$ is also harmonic in R and on C, and by assumption, $h - \Phi = 0$ everywhere on C. By (II) we thus have $h - \Phi = 0$ everywhere in R, and (III) is proved. ■

The last statement of Theorem 4 is very important. It means that a *harmonic function is uniquely determined in R by its values on the boundary of R*. Usually, $\Phi(x, y)$ is required to be harmonic in R and continuous on the boundary of R, that is,

$$\lim_{\substack{x \to x_0 \\ y \to y_0}} \Phi(x, y) = \Phi(x_0, y_0), \text{ where } (x_0, y_0) \text{ is on the boundary and } (x, y) \text{ is in } R.$$

Under these assumptions the maximum principle (I) is still applicable. The problem of determining $\Phi(x, y)$ when the boundary values are given is called the **Dirichlet problem** for the Laplace equation in two variables, as we know. From (III) we thus have, as a highlight of our discussion,

THEOREM 5

Uniqueness Theorem for the Dirichlet Problem

If for a given region and given boundary values the Dirichlet problem for the Laplace equation in two variables has a solution, the solution is unique.

PROBLEM SET 18.6

PROBLEMS RELATED TO THEOREMS 1 AND 2

1–4 Verify Theorem 1 for the given $F(z)$, z_0, and circle of radius 1.

1. $(z + 1)^3, \quad z_0 = \frac{5}{2}$
2. $2z^4, \quad z_0 = -2$
3. $(3z - 2)^2, \quad z_0 = 4$
4. $(z - 1)^{-2}, \quad z_0 = -1$
5. Integrate $|z|$ around the unit circle. Does the result contradict Theorem 1?
6. Derive the first statement in Theorem 2 from Poisson's integral formula.

7–9 Verify (3) in Theorem 2 for the given $\Phi(x, y)$, (x_0, y_0), and circle of radius 1.

7. $(x - 1)(y - 1), \quad (2, -2)$
8. $x^2 - y^2, \quad (3, 8)$
9. $x + y + xy, \quad (1, 1)$
10. Verify the calculations involving the inequalities in the proof of Theorem 3.
11. **CAS EXPERIMENT. Graphing Potentials.** Graph the potentials in Probs. 7 and 9 and for two other functions of your choice as surfaces over a rectangle in the xy-plane. Find the locations of the maxima and minima by inspecting these graphs.

12. TEAM PROJECT. Maximum Modulus of Analytic Functions. **(a)** Verify Theorem 3 for (i) $F(z) = z^2$ and the rectangle $1 \leqq x \leqq 5, 2 \leqq y \leqq 4$, (ii) $F(z) = \sin z$ and the unit disk, and (iii) $F(z) = e^z$ and any bounded domain.

(b) $F(z) = 1 + |z|$ is not zero in the disk $|z| \leqq 2$ and has a minimum at an interior point. Does this contradict Theorem 3?

(c) $F(x) = \sin x$ (x real) has a maximum 1 at $\pi/2$. Why can this not be a maximum of $|F(z)| = |\sin z|$ in a domain containing $z = \pi/2$?

(d) If $F(z)$ is analytic and not constant in the closed unit disk D: $|z| \leqq 1$ and $|F(z)| = c = \text{const}$ on the unit circle, show that $F(z)$ must have a zero in D.

13–17 MAXIMUM MODULUS

Find the location and size of the maximum of $|F(z)|$ in the unit disk $|z| \leqq 1$.

13. $F(z) = \cos z$

14. $F(z) = \exp z^2$

15. $F(z) = \sinh 2z$

16. $F(z) = az + b$ (a, b complex, $a \neq 0$)

17. $F(z) = 2z^2 - 2$

18. Verify the maximum principle for $\Phi(x, y) = e^x \sin y$ and the rectangle $a \leqq x \leqq b, 0 \leqq y \leqq 2\pi$.

19. Harmonic conjugate. Do Φ and a harmonic conjugate Ψ in a region R have their maximum at the same point of R?

20. Conformal mapping. Find the location (u_1, v_1) of the maximum of $\Phi^* = e^u \cos v$ in R^*: $|w| \leqq 1, v \geqq 0$, where $w = u + iv$. Find the region R that is mapped onto R^* by $w = f(z) = z^2$. Find the potential in R resulting from Φ^* and the location (x_1, y_1) of the maximum. Is (u_1, v_1) the image of (x_1, y_1)? If so, is this just by chance?

CHAPTER 18 REVIEW QUESTIONS AND PROBLEMS

1. Why can potential problems be modeled and solved by methods of complex analysis? For what dimensions?

2. What parts of complex analysis are mainly of interest to the engineer and physicist?

3. What is a harmonic function? A harmonic conjugate?

4. What areas of physics did we consider? Could you think of others?

5. Give some examples of potential problems considered in this chapter. Make a list of corresponding functions.

6. What does the *complex* potential give physically?

7. Write a short essay on the various assumptions made in fluid flow in this chapter.

8. Explain the use of conformal mapping in potential theory.

9. State the maximum modulus theorem and mean value theorems for harmonic functions.

10. State Poisson's integral formula. Derive it from Cauchy's formula.

11. Find the potential and the complex potential between the plates $y = x$ and $y = x + 10$ kept at 10 V and 110 V, respectively.

12. Find the potential and complex potential between the coaxial cylinders of axis 0 (hence the vertical axis in space) and radii $r_1 = 1$ cm, $r_2 = 10$ cm, kept at potential $U_1 = 200$ V and $U_2 = 2$ kV, respectively.

13. Do the task in Prob. 12 if $U_1 = 220$ V and the outer cylinder is grounded, $U_2 = 0$.

14. If plates at $x_1 = 1$ and $x_2 = 10$ are kept at potentials $U_1 = 200$ V, $U_2 = 2$ kV, is the potential at $x = 5$ larger or smaller than the potential at $r = 5$ in Prob. 12? No calculation. Give reason.

15. Make a list of important potential functions, with applications, from memory.

16. Find the equipotential lines of $F(z) = i \operatorname{Ln} z$.

17. Find the potential in the first quadrant of the xy-plane if the x-axis has potential 2 kV and the y-axis is grounded.

18. Find the potential in the angular region between the plates $\operatorname{Arg} z = \pi/6$ kept at 800 V and $\operatorname{Arg} z = \pi/3$ kept at 600 V.

19. Find the temperature T in the upper half-plane if, on the x-axis, $T = 30°$C for $x > 1$ and $-30°$C for $x < 1$.

20. Interpret Prob. 18 as an electrostatic problem. What are the lines of electric force?

21. Find the streamlines and the velocity for the complex potential $F(z) = (1 + i)z$. Describe the flow.

22. Describe the streamlines for $F(z) = \frac{1}{2}z^2 + z$.

23. Show that the isotherms of $F(z) = -iz^2 + z$ are hyperbolas.

24. State the theorem on the behavior of harmonic functions under conformal mapping. Verify it for $\Phi^* = e^u \sin v$ and $w = u + iv = z^2$.

25. Find V in Prob. 22 and verify that it gives vectors tangent to the streamlines.

SUMMARY OF CHAPTER 18
Complex Analysis and Potential Theory

Potential theory is the theory of solutions of **Laplace's equation**

$$\nabla^2 \Phi = 0. \tag{1}$$

Solutions whose second partial derivatives are ***continuous*** are called **harmonic functions**. Equation (1) is the most important PDE in physics, where it is of interest in two and three dimensions. It appears in electrostatics (Sec. 18.1), steady-state heat problems (Sec. 18.3), fluid flow (Sec. 18.4), gravity, etc. Whereas the three-dimensional case requires other methods (see Chap. 12), two-dimensional potential theory can be handled by complex analysis, since the real and imaginary parts of an analytic function are harmonic (Sec. 13.4). They remain harmonic under conformal mapping (Sec. 18.2), so that ***conformal mapping*** becomes a powerful tool in solving boundary value problems for (1), as is illustrated in this chapter. With a real potential Φ in (1) we can associate a **complex potential**

$$F(z) = \Phi + i\Psi \qquad \text{(Sec. 18.1).} \tag{2}$$

Then both families of curves $\Phi = \text{const}$ and $\Psi = \text{const}$ have a physical meaning. In electrostatics, they are equipotential lines and lines of electrical force (Sec. 18.1). In heat problems, they are isotherms (curves of constant temperature) and lines of heat flow (Sec. 18.3). In fluid flow, they are equipotential lines of the velocity potential and streamlines (Sec. 18.4).

For the disk, the solution of the Dirichlet problem is given by the ***Poisson formula*** (Sec. 18.5) or by a series that on the boundary circle becomes the Fourier series of the given boundary values (Sec. 18.5).

Harmonic functions, like analytic functions, have a number of general properties; particularly important are the ***mean value property*** and the ***maximum modulus property*** (Sec. 18.6), which implies the uniqueness of the solution of the Dirichlet problem (Theorem 5 in Sec. 18.6).

PART E

Numeric Analysis

Numeric analysis or briefly **numerics** continues to be one of the fastest growing areas of engineering mathematics. This is a natural trend with the ever greater availability of computing power and global Internet use. Indeed, good software implementation of numerical methods are readily available. Take a look at the ***updated*** list of ***Software*** starting on p. 788. It contains software for purchase (commercial software) and software for free download (public-domain software). For convenience, we provide Internet addresses and phone numbers. The software list includes computer algebra systems (CASs), such as ***Maple*** and ***Mathematica***, along with the *Maple Computer Guide*, 10th ed., and *Mathematica Computer Guide*, 10th ed., by E. Kreyszig and E. J. Norminton related to this text that teach you stepwise how to use these computer algebra systems and with complete engineering examples drawn from the text. Furthermore, there is scientific software, such as ***IMSL***, ***LAPACK*** (free download), and scientific calculators with graphic capabilities such as ***TI-Nspire***. Note that, although we have listed frequently used quality software, this list is by no means complete.

In your career as an engineer, appplied mathematician, or scientist you are likely to use commercially available software or proprietary software, owned by the company you work for, that uses numeric methods to solve engineering problems, such as modeling chemical or biological processes, planning ecologically sound heating systems, or computing trajectories of spacecraft or satellites. For example, one of the collaborators of this book (Herbert Kreyszig) used proprietary software to determine the value of bonds, which amounted to solving higher degree polynomial equations, using numeric methods discussed in Sec. 19.2.

However, the availability of quality software does not alleviate your effort and responsibility to first ***understand*** *these numerical methods.* Your effort will pay off because, with your mathematical expertise in numerics, you will be able to plan your solution approach, judiciously select and use the appropriate software, judge the quality of software, and, perhaps, even write your own numerics software.

Numerics extends your ability to solve problems that are either difficult or impossible to solve analytically. For example, certain integrals such as error function [see App. 3, formula (35)] or large eigenvalue problems that generate high-degree characteristic polynomials cannot be solved analytically. Numerics is also used to construct approximating polynomials through data points that were obtained from some experiments.

Part E is designed to give you a solid background in numerics. We present many numeric methods as **algorithms**, which give these methods in detailed steps suitable for software implementation on your computer, CAS, or programmable calculator. The first chapter, Chap. 19, covers three main areas. These are general numerics (floating point, rounding errors, etc.), solving equations of the form $f(x) = 0$ (using Newton's method and other methods), interpolation along with methods of numeric integration that make use of it, and differentiation.

Chapter 20 covers the essentials of numeric linear algebra. The chapter breaks into two parts: solving linear systems of equations by methods of Gauss, Doolittle, Cholesky, etc. and solving eigenvalue problems numerically. Chapter 21 again has two themes: solving ordinary differential equations and systems of ordinary differential equations as well as solving partial differential equations.

Numerics is a very active area of research as new methods are invented, existing methods improved and adapted, and old methods—impractical in precomputer times—are rediscovered. A main goal in these activities is the development of well-structured software. And in large-scale work—millions of equations or steps of iterations—even small algorithmic improvements may have a large significant effect on computing time, storage demand, accuracy, and stability.

Remark on Software Use. Part E is designed in such a way as to *allow compelete flexibility on the use of CASs, software, or graphing calculators.* The computational requirements range from very little use to heavy use. The choice of computer use is at the discretion of the professor. The material and problem sets (except where clearly indicated such as in CAS Projects, CAS Problems, or CAS Experiments, which can be omitted without loss of continuity) do not require the use of a CAS or software. A scientific calculator perhaps with graphing capabilities is all that is required.

Software

See also http://www.wiley.com/college/kreyszig/

The following list will help you if you wish to find software. You may also obtain information on known and new software from websites such as Dr. Dobb's Portal, from articles published by the *American Mathematical Society* (see also its website at www.ams.org), the *Society for Industrial and Applied Mathematics* (SIAM, at www.siam.org), the *Association for Computing Machinery* (ACM, at www.acm.org), or the *Institute of Electrical and Electronics Engineers* (IEEE, at www.ieee.org). Consult also your library, computer science department, or mathematics department.

TI-Nspire. Includes TI-Nspire CAS and programmable graphic calculators. Texas Instruments, Inc., Dallas, TX. Telephone: 1-800-842-2737 or (972) 917-8324; website at www.education.ti.com.

EISPACK. See LAPACK.

GAMS (Guide to Available Mathematical Software). Website at http://gams.nist.gov. Online cross-index of software development by NIST.

IMSL (International Mathematical and Statistical Library). Visual Numerics, Inc., Houston, TX. Telephone: 1-800-222-4675 or (713) 784-3131; website at www.vni.com. Mathematical and statistical FORTRAN routines with graphics.

LAPACK. FORTRAN 77 routines for linear algebra. This software package supersedes LINPACK and EISPACK. You can download the routines from www.netlib.org/lapack. The LAPACK User's Guide is available at www.netlib.org.

LINPACK see LAPACK

Maple. Waterloo Maple, Inc., Waterloo, ON, Canada. Telephone: 1-800-267-6583 or (519) 747-2373; website at www.maplesoft.com.

Maple Computer Guide. For Advanced Engineering Mathematics, 10th edition. By E. Kreyszig and E. J. Norminton. John Wiley and Sons, Inc., Hoboken, NJ. Telephone: 1-800-225-5945 or (201) 748-6000.

Mathcad. Parametric Technology Corp. (PTC), Needham, MA. Website at www.ptc.com.

Mathematica. Wolfram Research, Inc., Champaign, IL. Telephone: 1-800-965-3726 or (217) 398-0700; website at www.wolfram.com.

Mathematica Computer Guide. For Advanced Engineering Mathematics, 10th edition. By E. Kreyszig and E. J. Norminton. John Wiley and Sons, Inc., Hoboken, NJ. Telephone: 1-800-225-5945 or (201) 748-6000.

Matlab. The MathWorks, Inc., Natick, MA. Telephone: (508) 647-7000; website at www.mathworks.com.

NAG. Numerical Algorithms Group, Inc., Lisle, IL. Telephone: (630) 971-2337; website at www.nag.com. Numeric routines in FORTRAN 77, FORTRAN 90, and C.

NETLIB. Extensive library of public-domain software. See at www.netlib.org.

NIST. National Institute of Standards and Technology, Gaithersburg, MD. Telephone: (301) 975-6478; website at www.nist.gov. For Mathematical and Computational Science Division telephone: (301) 975-3800. See also http://math.nist.gov.

Numerical Recipes. Cambridge University Press, New York, NY. Telephone: 1-800-221-4512 or (212) 924-3900; website at www.cambridge.org/us. Book, 3rd ed. (in C++) see App. 1, Ref. [E25]; source code on CD ROM in C++, which also contains old source code (but not text) for (out of print) 2nd ed. C, FORTRAN 77, FORTRAN 90 as well as source code for (out of print) 1st ed. To order, call office at West Nyack, NY, at 1-800-872-7423 or (845) 353-7500 or online at www.nr.com.

FURTHER SOFTWARE IN STATISTICS. See Part G.

CHAPTER 19

Numerics in General

Numeric analysis or briefly numerics has a distinct flavor that is different from basic calculus, from solving ODEs algebraically, or from other (nonnumeric) areas. Whereas in calculus and in ODEs there were very few choices on how to solve the problem and your answer was an algebraic answer, in numerics you have many more choices and your answers are given as tables of values (numbers) or graphs. You have to make judicous choices as to what numeric method or algorithm you want to use, how accurate you need your result to be, with what value (starting value) do you want to begin your computation, and others. This chapter is designed to provide a good transition from the algebraic type of mathematics to the numeric type of mathematics.

We begin with the general concepts such as floating point, roundoff errors, and general numeric errors and their propagation. This is followed in Sec. 19.2 by the important topic of solving equations of the type $f(x) = 0$ by various numeric methods, including the famous Newton method. Section 19.3 introduces interpolation methods. These are methods that construct new (unknown) function values from known function values. The knowledge gained in Sec. 19.3 is applied to spline interpolation (Sec. 19.4) and is useful for understanding numeric integration and differentiation covered in the last section.

Numerics provides an invaluable extension to the knowledge base of the problem-solving engineer. Many problems have no solution formula (think of a complicated integral or a polynomial of high degree or the interpolation of values obtained by measurements). In other cases a complicated solution formula may exist but may be practically useless. It is for these kinds of problems that a numerical method may generate a good answer. Thus, it is very important that the applied mathematician, engineer, physicist, or scientist becomes familiar with the essentials of numerics and its ideas, such as estimation of errors, order of convergence, numerical methods expressed in algorithms, and is also informed about the important numeric methods.

Prerequisite: Elementary calculus.
References and Answers to Problems: App. 1 Part E, App. 2.

19.1 Introduction

As an engineer or physicist you may deal with problems in elasticity and need to solve an equation such as $x \cosh x = 1$ or a more difficult problem of finding the roots of a higher order polynomial. Or you encounter an integral such as

$$\int_0^1 \exp(-x^2)\, dx$$

[see App. 3, formula (35)] that you cannot solve by elementary calculus. Such problems, which are difficult or impossible to solve algebraically, arise frequently in applications. They call for **numeric methods**, that is, systematic methods that are suitable for solving, numerically, the problems on computers or calculators. Such solutions result in tables of numbers, graphical representation (figures), or both. Typical numeric methods are iterative in nature and, for a well-choosen problem and a good starting value, will frequently converge to a desired answer. The evolution from a given problem that you observed in an experimental lab or in an industrial setting (in engineering, physics, biology, chemistry, economics, etc.) to an approximation suitable for numerics to a final answer usually requires the following steps.

1. **Modeling.** We set up a mathematical model of our problem, such as an integral, a system of equations, or a differential equation.
2. **Choosing a numeric method** and parameters (e.g., step size), perhaps with a preliminary error estimation.
3. **Programming.** We use the algorithm to write a corresponding program in a CAS, such as Maple, Mathematica, Matlab, or Mathcad, or, say, in Java, C or C^{++}, or FORTRAN, selecting suitable routines from a software system as needed.
4. **Doing the computation.**
5. **Interpreting the results** in physical or other terms, also deciding to rerun if further results are needed.

Steps 1 and 2 are related. A slight change of the model may often admit of a more efficient method. To choose methods, we must first get to know them. Chapters 19–21 contain efficient algorithms for the most important classes of problems occurring frequently in practice.

In Step 3 the program consists of the given data and a sequence of instructions to be executed by the computer in a certain order for producing the answer in numeric or graphic form.

To create a good understanding of the nature of numeric work, we continue in this section with some simple general remarks.

Floating-Point Form of Numbers

We know that in decimal notation, every real number is represented by a finite or an infinite sequence of decimal digits. Now most computers have two ways of representing numbers, called *fixed point* and *floating point*. In a **fixed-point** system all numbers are given with a fixed number of decimals after the decimal point; for example, numbers given with 3 decimals are 62.358, 0.014, 1.000. In a text we would write, say, 3 decimals as 3D. Fixed-point representations are impractical in most scientific computations because of their limited range (explain!) and will not concern us.

In a **floating-point** system we write, for instance,

$$0.6247 \cdot 10^{3}, \qquad 0.1735 \cdot 10^{-13}, \qquad -0.2000 \cdot 10^{-1}$$

or sometimes also

$$6.247 \cdot 10^{2}, \qquad 1.735 \cdot 10^{-14}, \qquad -2.000 \cdot 10^{-2}.$$

We see that in this system the number of significant digits is kept fixed, whereas the decimal point is "floating." Here, a **significant digit** of a number c is any given digit of c, except

possibly for zeros to the left of the first nonzero digit; these zeros serve only to fix the position of the decimal point. (Thus any other zero is a significant digit of c.) For instance,

$$13600, \quad 1.3600, \quad 0.0013600$$

all have 5 significant digits. In a text we indicate, say, 5 significant digits, by 5S.

The use of exponents permits us to represent very large and very small numbers. Indeed, theoretically any nonzero number a can be written as

$$(1) \qquad a = \pm m \cdot 10^n, \qquad 0.1 \leqq |m| < 1, \qquad n \text{ integer.}$$

On modern computers, which use binary (base 2) numbers, m is limited to k binary digits (e.g., $k = 8$) and n is limited (see below), giving representations (for finitely many numbers only!)

$$(2) \qquad \bar{a} = \pm \bar{m} \cdot 2^n, \qquad \bar{m} = 0.d_1 d_2 \cdots d_k, \qquad d_1 > 0.$$

These numbers $\bar{a}$ are called *k-digit binary* **machine numbers**. Their fractional part m (or $\bar{m}$) is called the *mantissa*. This is not identical with "mantissa" as used for logarithms. n is called the *exponent* of $\bar{a}$.

It is important to realize that there are only finitely many machine numbers and that they become less and less "dense" with increasing a. For instance, there are as many numbers between 2 and 4 as there are between 1024 and 2048. Why?

The smallest positive machine number *eps* with $1 + \text{eps} > 1$ is called the *machine accuracy*. It is important to realize that there are no numbers in the intervals $[1, 1 + \text{eps}]$, $[2, 2 + 2 \cdot \text{eps}], \cdots, [1024, 1024 + 1024 \cdot \text{eps}], \cdots$. This means that, if the mathematical answer to a computation would be $1024 + 1024 \cdot \text{eps}/2$, the computer result will be *either* 1024 or $1024 \cdot \text{eps}$ so it is impossible to achieve greater accuracy.

Underflow and Overflow. The range of exponents that a typical computer can handle is very large. The IEEE (Institute of Electrical and Electronic Engineers) floating-point standard for **single precision** is from 2^{-126} to 2^{128} (1.175×10^{-38} to 3.403×10^{38}) and for **double precision** it is from 2^{-1022} to 2^{1024} (2.225×10^{-308} to 1.798×10^{308}).

As a minor technicality, to avoid storing a minus in the exponent, the ranges are shifted from $[-126, 128]$ by adding 126 (for double precision 1022). Note that shifted exponents of 255 and 1047 are used for some special cases such as representing infinity.

If, in a computation a number outside that range occurs, this is called **underflow** when the number is smaller and **overflow** when it is larger. In the case of underflow, the result is usually set to zero and computation continues. Overflow might cause the computer to halt. Standard codes (by IMSL, NAG, etc.) are written to avoid overflow. Error messages on overflow may then indicate programming errors (incorrect input data, etc.). From here on, we will be discussing the decimal results that we obtain from our computations.

Roundoff

An error is caused by **chopping** (= discarding all digits from some decimal on) or **rounding**. This error is called **roundoff error**, regardless of whether we chop or round. The rule for rounding off a number to k decimals is as follows. (The rule for rounding off to k significant digits is the same, with "decimal" replaced by "significant digit.")

Roundoff Rule. To round a number x to k decimals, and $5 \cdot 10^{-(k+1)}$ to x and chop the digits after the $(k + 1)$st digit.

EXAMPLE 1 **Roundoff Rule**

Round the number 1.23454621 to **(a)** 2 decimals, **(b)** 3 decimals, **(c)** 4 decimals, **(d)** 5 decimals, and **(e)** 6 decimals.

Solution. **(a)** For 2 decimals we add $5 \cdot 10^{-(k+1)} = 5 \cdot 10^{-3} = 0.005$ to the given number, that is, $1.2345621 + 0.005 = 1.23\,954621$. Then we chop off the digits "954621" after the space or equivalently $1.23954621 - 0.00954621 = 1.23$.
(b) $1.23454621 + 0.0005 = 1.235\,04621$, so that for 3 decimals we get 1.234.
(c) 1.23459621 after chopping give us 1.2345 (4 decimals).
(d) 1.23455121 yields 1.23455 (5 decimals).
(e) 1.23454671 yields 1.234546 (6 decimals).
Can you round the number to 7 decimals? ■

Chopping is not recommended because the corresponding error can be larger than that in rounding. (Nevertheless, some computers use it because it is simpler and faster. On the other hand, some computers and calculators improve accuracy of results by doing intermediate calculations using one or more extra digits, called *guarding digits.*)

Error in Rounding. Let $\bar{a} = \text{fl}(a)$ in (2) be the floating-point computer approximation of a in (1) obtained by rounding, where fl suggests **floating**. Then the roundoff rule gives (by dropping exponents) $|m - \bar{m}| \leqq \frac{1}{2} \cdot 10^{-k}$. Since $|m| \geqq 0.1$, this implies (when $a \neq 0$)

$$\left|\frac{a - \bar{a}}{a}\right| \approx \left|\frac{m - \bar{m}}{m}\right| \leqq \frac{1}{2} \cdot 10^{1-k}. \tag{3}$$

The right side $u = \frac{1}{2} \cdot 10^{1-k}$ is called the **rounding unit**. If we write $\bar{a} = a(1 + \delta)$, we have by algebra $(\bar{a} - a)/a = \delta$, hence $|\delta| \leqq u$ by (3). *This shows that the rounding unit u is an error bound in rounding.*

Rounding errors may ruin a computation completely, even a small computation. In general, these errors become the more dangerous the more arithmetic operations (perhaps several millions!) we have to perform. It is therefore important to analyze computational programs for expected rounding errors and to find an arrangement of the computations such that the effect of rounding errors is as small as possible.

As mentioned, the arithmetic in a computer is not exact and causes further errors; however, these will not be relevant to our discussion.

Accuracy in Tables. Although available software has rendered various tables of function values superfluous, some tables (of higher functions, of coefficients of integration formulas, etc.) will still remain in occasional use. If a table shows k significant digits, it is conventionally assumed that any value $\tilde{a}$ in the table deviates from the exact value a by at most $\pm\frac{1}{2}$ unit of the kth digit.

Loss of Significant Digits

This means that a result of a calculation has fewer correct digits than the numbers from which it was obtained. This happens if we subtract two numbers of about the same size, for example, $0.1439 - 0.1426$ ("subtractive cancellation"). It may occur in simple problems, but it can be avoided in most cases by simple changes of the algorithm—if one is aware of it! Let us illustrate this with the following basic problem.

EXAMPLE 2 **Quadratic Equation. Loss of Significant Digits**

Find the roots of the equation

$$x^2 + 40x + 2 = 0,$$

using 4 significant digits (abbreviated 4S) in the computation.

Solution. A formula for the roots x_1, x_2 of a quadratic equation $ax^2 + bx + c = 0$ is

$$(4) \qquad x_1 = \frac{1}{2a}\left(-b + \sqrt{b^2 - 4ac}\right), \qquad x_2 = \frac{1}{2a}\left(-b - \sqrt{b^2 - 4ac}\right).$$

Furthermore, since $x_1 x_2 = c/a$, another formula for those roots

$$(5) \qquad x_1 = \frac{c}{ax_2}, \qquad x_2 \text{ as in (4).}$$

We see that this avoids cancellation in x_1 for positive b.

If $b < 0$, calculate x_1 from (4) and then $x_2 = c/(ax_1)$.

For $x^2 + 40x + 2 = 0$ we obtain from (4) $x = -20 \pm \sqrt{398} = -20 \pm 19.95$, hence $x_2 = -20.00 - 19.95$, involving no difficulty, and $x_1 = -20.00 + 19.95 = -0.05$, a poor value involving loss of digits by subtractive cancellation.

In contrast, (5) gives $x_1 = 2.000/(-39.95) = -0.05006$, the absolute value of the error being less than one unit of the last digit, as a computation with more digits shows. The 10S-value is -0.05006265674. ■

Errors of Numeric Results

Final results of computations of unknown quantities generally are **approximations**; that is, they are not exact but involve errors. Such an error may result from a combination of the following effects. **Roundoff errors** result from rounding, as discussed above. **Experimental errors** are errors of given data (probably arising from measurements). **Truncating errors** result from truncating (prematurely breaking off), for instance, if we replace a Taylor series with the sum of its first few terms. These errors depend on the computational method used and must be dealt with individually for each method. [“Truncating” is sometimes used as a term for chopping off (see before), a terminology that is not recommended.]

Formulas for Errors. If $\tilde{a}$ is an approximate value of a quantity whose exact value is a, we call the difference

$$\textbf{(6)} \qquad \epsilon = a - \tilde{a}$$

the **error** of $\tilde{a}$. Hence

$$\textbf{(6*)} \qquad a = \tilde{a} + \epsilon, \qquad \text{True value} = \text{Approximation} + \text{Error}.$$

For instance, if $\tilde{a} = 10.5$ is an approximation of $a = 10.2$, its error is $\epsilon = -0.3$. The error of an approximation $\tilde{a} = 1.60$ of $a = 1.82$ is $\epsilon = 0.22$.

CAUTION! In the literature $|a - \tilde{a}|$ (“absolute error”) or $\tilde{a} - a$ are sometimes also used as definitions of error.

The **relative error** ϵ_r of $\tilde{a}$ is defined by

$$(7) \qquad \epsilon_r = \frac{\epsilon}{a} = \frac{a - \tilde{a}}{a} = \frac{\text{Error}}{\text{True value}} \qquad (a \neq 0).$$

This looks useless because a is unknown. But if $|\epsilon|$ is much less than $|\tilde{a}|$, then we can use $\tilde{a}$ instead of a and get

$$(7') \qquad \epsilon_r \approx \frac{\epsilon}{\tilde{a}}.$$

This still looks problematic because ϵ is unknown—if it were known, we could get $a = \tilde{a} + \epsilon$ from (6) and we would be done. But what one often can obtain in practice is an **error bound** for $\tilde{a}$, that is, a number β such that

$$|\epsilon| \leqq \beta, \qquad \text{hence} \qquad |a - \tilde{a}| \leqq \beta.$$

This tells us how far away from our computed $\tilde{a}$ the unknown a can at most lie. Similarly, for the relative error, an error bound is a number β_r such that

$$|\epsilon_r| \leqq \beta_r, \qquad \text{hence} \qquad \left|\frac{a - \tilde{a}}{a}\right| \leqq \beta_r.$$

Error Propagation

This is an important matter. It refers to how errors at the beginning and in later steps (roundoff, for example) propagate into the computation and affect accuracy, sometimes very drastically. We state here what happens to error bounds. Namely, bounds for the *error* add under addition and subtraction, whereas bounds for the *relative error* add under multiplication and division. You do well to keep this in mind.

THEOREM 1

Error Propagation

(a) *In addition and subtraction, a bound for the* ***error*** *of the results is given by the sum of the error bounds for the terms.*

(b) *In multiplication and division, an error bound for the* ***relative error*** *of the results is given (approximately) by the sum of the bounds for the relative errors of the given numbers.*

PROOF **(a)** We use the notations $x = \tilde{x} + \epsilon_x$, $y = \tilde{y} + \epsilon_y$, $|\epsilon_x| \leqq \beta_x$, $|\epsilon_y| \leqq \beta_y$. Then for the error ϵ of the *difference* we obtain

$$\begin{aligned} |\epsilon| &= |x - y - (\tilde{x} - \tilde{y})| \\ &= |x - \tilde{x} - (y - \tilde{y})| \\ &= |\epsilon_x - \epsilon_y| \leqq |\epsilon_x| + |\epsilon_y| \leqq \beta_x + \beta_y. \end{aligned}$$

The proof for the *sum* is similar and is left to the student.

(b) For the relative error ϵ_r of $\tilde{x}\tilde{y}$ we get from the relative errors ϵ_{rx} and ϵ_{ry} of $\tilde{x}, \tilde{y}$ and bounds β_{rx}, β_{ry}

$$\begin{aligned} |\epsilon_r| &= \left|\frac{xy - \tilde{x}\tilde{y}}{xy}\right| = \left|\frac{xy - (x - \epsilon_x)(y - \epsilon_y)}{xy}\right| = \left|\frac{\epsilon_x y + \epsilon_y x - \epsilon_x \epsilon_y}{xy}\right| \\ &\approx \left|\frac{\epsilon_x y + \epsilon_y x}{xy}\right| \leqq \left|\frac{\epsilon_x}{x}\right| + \left|\frac{\epsilon_y}{y}\right| = |\epsilon_{rx}| + |\epsilon_{ry}| \leqq \beta_{rx} + \beta_{ry}. \end{aligned}$$

This proof shows what "approximately" means: we neglected $\epsilon_x \epsilon_y$ as small in absolute value compared to $|\epsilon_x|$ and $|\epsilon_y|$. The proof for the quotient is similar but slightly more tricky (see Prob. 13). ■

Basic Error Principle

Every numeric method should be accompanied by an error estimate. If such a formula is lacking, is extremely complicated, or is impractical because it involves information (for instance, on derivatives) that is not available, the following may help.

Error Estimation by Comparison. *Do a calculation twice with different accuracy. Regard the difference $\tilde{a}_2 - \tilde{a}_1$ of the results $\tilde{a}_1, \tilde{a}_2$ as a (perhaps crude) estimate of the error ϵ_1 of the inferior result $\tilde{a}_1$.* Indeed, $\tilde{a}_1 + \epsilon_1 = \tilde{a}_2 + \epsilon_2$ by formula (4*). This implies $\tilde{a}_2 - \tilde{a}_1 = \epsilon_1 - \epsilon_2 \approx \epsilon_1$ because $\tilde{a}_2$ is generally more accurate than $\tilde{a}_1$, so that $|\epsilon_2|$ is small compared to $|\epsilon_1|$.

Algorithm. Stability

Numeric methods can be formulated as algorithms. An **algorithm** is a step-by-step procedure that states a numeric method in a form (a "**pseudocode**") understandable to humans. (See Table 19.1 to see what an algorithm looks like.) The algorithm is then used to write a program in a programming language that the computer can understand so that it can execute the numeric method. Important algorithms follow in the next sections. For routine tasks your CAS or some other software system may contain programs that you can use or include as parts of larger programs of your own.

Stability. To be useful, an algorithm should be **stable**; that is, small changes in the initial data should cause only small changes in the final results. However, if small changes in the initial data can produce large changes in the final results, we call the algorithm **unstable**.

This "*numeric instability*," which in most cases can be avoided by choosing a better algorithm, must be distinguished from "*mathematical instability*" of a problem, which is called "*ill-conditioning*," a concept we discuss in the next section.

Some algorithms are stable only for certain initial data, so that one must be careful in such a case.

PROBLEM SET 19.1

1. **Floating point.** Write 84.175, −528.685, 0.000924138, and −362005 in floating-point form, rounded to 5S (5 significant digits).
2. Write −76.437125, 60100, and −0.00001 in floating-point form, rounded to 4S.
3. **Small differences of large numbers** may be particularly strongly affected by rounding errors. Illustrate this by computing $0.81534/(35 \cdot 724 - 35.596)$ as given with 5S, then rounding stepwise to 4S, 3S, and 2S, where "stepwise" means round the rounded numbers, not the given ones.
4. **Order of terms**, in adding with a fixed number of digits, will generally affect the sum. Give an example. Find empirically a rule for the best order.
5. **Rounding and adding.** Let $a_1, \cdots, a_n$ be numbers with a_j correctly rounded to S_j digits. In calculating the sum $a_1 + \cdots + a_n$, retaining $S = \min S_j$ significant digits, is it essential that we first add and then round the result or that we first round each number to S significant digits and then add?
6. **Nested form.** Evaluate

$$f(x) = x^3 - 7.5x^2 + 11.2x + 2.8$$
$$= ((x - 7.5)x + 11.2)x + 2.8$$

at $x = 3.94$ using 3S arithmetic and rounding, in both of the given forms. The latter, called the *nested form*, is usually preferable since it minimizes the number of operations and thus the effect of rounding.

7. **Quadratic equation.** Solve $x^2 - 30x + 1 = 0$ by (4) and by (5), using 6S in the computation. Compare and comment.

8. Solve $x^2 - 40x + 2 = 0$, using 4S-computation.

9. Do the computations in Prob. 7 with 4S and 2S.

10. **Instability.** For small $|a|$ the equation $(x - k)^2 = a$ has nearly a double root. Why do these roots show instability?

11. **Theorems on errors.** Prove Theorem 1(a) for addition.

12. **Overflow and underflow** can sometimes be avoided by simple changes in a formula. Explain this in terms of $\sqrt{x^2 + y^2} = x\sqrt{1 + (y/x)^2}$ with $x^2 \geqq y^2$ and x so large that x^2 would cause overflow. Invent examples of your own.

13. **Division.** Prove Theorem 1(b) for division.

14. **Loss of digits. Square root.** Compute $\sqrt{x^2 + 4} - 2$ with 6S arithmetic for $x = 0.001$ **(a)** as given and **(b)** from $x^2/(\sqrt{x^2 + 4} + 2)$ (derive!).

15. **Logarithm.** Compute $\ln a - \ln b$ with 6S arithmetic for $a = 4.00000$ and $b = 3.99900$ **(a)** as given and **(b)** from $\ln(a/b)$.

16. **Cosine.** Compute $1 - \cos x$ with 6S arithmetic for $x = 0.02$ **(a)** as given and **(b)** by $2 \sin^2 \frac{1}{2}x$ (derive!).

17. Discuss the numeric use of (12) in App. A3.1 for $\cos v - \cos u$ when $u \approx v$.

18. **Quotient near 0/0. (a)** Compute $(1 - \cos x)/\sin x$ with 6S arithmetic for $x = 0.005$. **(b)** Looking at Prob. 16, find a much better formula.

19. **Exponential function.** Calculate $1/e = 0.367879$ (6S) from the partial sums of 5–10 terms of the Maclaurin series **(a)** of e^{-x} with $x = 1$, **(b)** of e^x with $x = 1$ and then taking the reciprocal. Which is more accurate?

20. Compute e^{-10} with 6S arithmetic in two ways (as in Prob. 19).

21. **Binary conversion.** Show that

$$\begin{aligned} 23 &= 20 \cdot 10^1 + 3 \cdot 10^0 = 16 + 4 + 2 + 1 \\ &= 2^4 + 2^2 + 2^1 + 2^0 = (1 \quad 0 \quad 1 \quad 1 \quad 1.)_2 \end{aligned}$$

can be obtained by the division algorithm

$$\begin{array}{lll} 2\,\underline{|23} & \text{Remainder} & 1 = c_0 \\ 2\,\underline{|11} & & 1 = c_1 \\ 2\,\underline{|5\ } & & 1 = c_2 \\ 2\,\underline{|2\ } & & 0 = c_3 \\ \ \ 0 & & 1 = c_4 \end{array}$$

22. Convert $(0.59375)_{10}$ to $(0.10011)_2$ by successive multiplication by 2 and dropping (removing) the integer parts, which give the binary digits $c_1, c_2, \cdots$:

$$\begin{array}{rl} & 0\ .59375 \cdot 2 \\ c_1 = & \boxed{1}\ .1875 \cdot 2 \\ c_2 = & \boxed{0}\ .375 \cdot 2 \\ c_3 = & \boxed{0}\ .75 \cdot 2 \\ c_4 = & \boxed{1}\ .5 \cdot 2 \\ c_5 = & \boxed{1}\ .0 \end{array}$$

23. Show that 0.1 is not a binary machine number.

24. Prove that any binary machine number has a finite decimal representation. Is the converse true?

25. **CAS EXPERIMENT. Approximations.** Obtain $x = 0.1 = \frac{3}{2} \sum_{m=1}^{\infty} 2^{-4m}$ from Prob. 23. Which machine number (partial sum) S_n will first have the value 0.1 to 30 decimal digits?

26. **CAS EXPERIMENT. Integration from Calculus.** Integrating by parts, show that $I_n = \int_0^1 e^x x^n \, dx = e - nI_{n-1}$, $I_0 = e - 1$. **(a)** Compute I_n, $n = 0, \cdots$, using 4S arithmetic, obtaining $I_8 = -3.906$. Why is this nonsense? Why is the error so large?

    (b) Experiment in (a) with the number of digits $k > 4$. As you increase k, will the first negative value $n = N$ occur earlier or later? Find an empirical formula for $N = N(k)$.

27. **Backward Recursion. In Prob. 26.** Using $e^x < e$ $(0 < x < 1)$, conclude that $|I_n| \leqq e/(n + 1) \rightarrow 0$ as $n \rightarrow \infty$. Solve the iteration formula for $I_{n-1} = (e - I_n)/n$, start from $I_{15} \approx 0$ and compute 4S values of $I_{14}, I_{13}, \cdots, I_1$.

28. **Harmonic series.** $1 + \frac{1}{2} + \frac{1}{3} + \cdots$ diverges. Is the same true for the corresponding series of computer numbers?

29. **Approximations of π = 3.14159265358979** $\cdots$ are 22/7 and 355/113. Determine the corresponding errors and relative errors to 3 significant digits.

30. **Compute π by Machin's approximation** $16 \arctan(\frac{1}{5}) - 4 \arctan(\frac{1}{239})$ to 10S (which are correct). [In 1986, D. H. Bailey (NASA Ames Research Center, Moffett Field, CA 94035) computed almost 30 million decimals of π on a CRAY-2 in less than 30 hrs. The race for more and more decimals is continuing. See the Internet under pi.]

19.2 Solution of Equations by Iteration

For each of the remaining sections of this chapter, we select basic kinds of problems and discuss numeric methods on how to solve them. The reader will learn about a variety of important problems and become familiar with ways of thinking in numerical analysis.

Perhaps the easiest conceptual problem is to find solutions of a single equation

$$f(x) = 0, \tag{1}$$

where f is a given function. A **solution** of (1) is a number $x = s$ such that $f(s) = 0$. Here, s suggests "solution," but we shall also use other letters.

It is interesting to note that the task of solving (1) is a question made for numeric algorithms, as in general there are no direct formulas, except in a few simple cases.

Examples of single equations are $x^3 + x = 1$, $\sin x = 0.5x$, $\tan x = x$, $\cosh x = \sec x$, $\cosh x \cos x = -1$, which can all be written in the form of (1). The first of the five equations is an **algebraic equation** because the corresponding f is a polynomial. In this case the solutions are called **roots** of the equation and the solution process is called *finding roots*. The other equations are **transcendental equations** because they involve transcendental functions.

There are a very large number of applications in engineering, where we have to solve a single equation (1). You have seen such applications when solving characteristic equations in Chaps. 2, 4, and 8; partial fractions in Chap. 6; residue integration in Chap. 16, finding eigenvalues in Chap. 12, and finding zeros of Bessel functions, also in Chap. 12. Moreover, methods of finding roots are very important in areas outside of classical engineering. For example, in finance, the problem of determining how much a bond is worth amounts to solving an algebraic equation.

To solve (1) when there is no formula for the exact solution available, we can use an approximation method, such as an **iteration method**. This is a method in which we start from an initial guess x_0 (which may be poor) and compute step by step (in general better and better) approximations $x_1, x_2, \cdots$ of an unknown solution of (1). We discuss three such methods that are of particular practical importance and mention two others in the problem set.

It is very important that the reader understand these methods and their underlying ideas. The reader will then be able to select judiciously the appropriate software from among different software packages that employ variations of such methods and not just treat the software programs as "black boxes."

In general, iteration methods are easy to program because the computational operations are the same in each step—just the data change from step to step—and, more importantly, if in a concrete case a method converges, it is stable in general (see Sec. 19.1).

Fixed-Point Iteration for Solving Equations $f(x) = 0$

Note: Our present use of the word "fixed point" has absolutely nothing to do with that in the last section.

By some *algebraic steps* we transform (1) into the form

$$x = g(x). \tag{2}$$

Then we choose an x_0 and compute $x_1 = g(x_0)$, $x_2 = g(x_1)$, and in general

$$x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots). \tag{3}$$

A solution of (2) is called a **fixed point** of g, motivating the name of the method. This is a solution of (1), since from $x = g(x)$ we can return to the original form $f(x) = 0$. From (1) we may get several different forms of (2). The behavior of corresponding iterative sequences $x_0, x_1, \cdots$ may differ, in particular, with respect to their speed of convergence. Indeed, some of them may not converge at all. Let us illustrate these facts with a simple example.

EXAMPLE 1 An Iteration Process (Fixed-Point Iteration)

Set up an iteration process for the equation $f(x) = x^2 - 3x + 1 = 0$. Since we know the solutions

$$x = 1.5 \pm \sqrt{1.25}, \qquad \text{thus} \qquad 2.618034 \qquad \text{and} \qquad 0.381966,$$

we can watch the behavior of the error as the iteration proceeds.

Solution. The equation may be written

$$\text{(4a)} \qquad x = g_1(x) = \tfrac{1}{3}(x^2 + 1), \qquad \text{thus} \qquad x_{n+1} = \tfrac{1}{3}(x_n^2 + 1).$$

If we choose $x_0 = 1$, we obtain the sequence (Fig. 426a; computed with 6S and then rounded)

$$x_0 = 1.000, \quad x_1 = 0.667, \quad x_2 = 0.481, \quad x_3 = 0.411, \quad x_4 = 0.390, \cdots$$

which seems to approach the smaller solution. If we choose $x_0 = 2$, the situation is similar. If we choose $x_0 = 3$, we obtain the sequence (Fig. 426a, upper part)

$$x_0 = 3.000, \quad x_1 = 3.333, \quad x_2 = 4.037, \quad x_3 = 5.766, \quad x_4 = 11.415, \cdots$$

which diverges.

Our equation may also be written (divide by x)

$$\text{(4b)} \qquad x = g_2(x) = 3 - \frac{1}{x}, \qquad \text{thus} \qquad x_{n+1} = 3 - \frac{1}{x_n},$$

and if we choose $x_0 = 1$, we obtain the sequence (Fig. 426b)

$$x_0 = 1.000, \quad x_1 = 2.000, \quad x_2 = 2.500, \quad x_3 = 2.600, \quad x_4 = 2.615, \cdots$$

which seems to approach the larger solution. Similarly, if we choose $x_0 = 3$, we obtain the sequence (Fig. 426b)

$$x_0 = 3.000, \quad x_1 = 2.667, \quad x_2 = 2.625, \quad x_3 = 2.619, \quad x_4 = 2.618, \cdots.$$

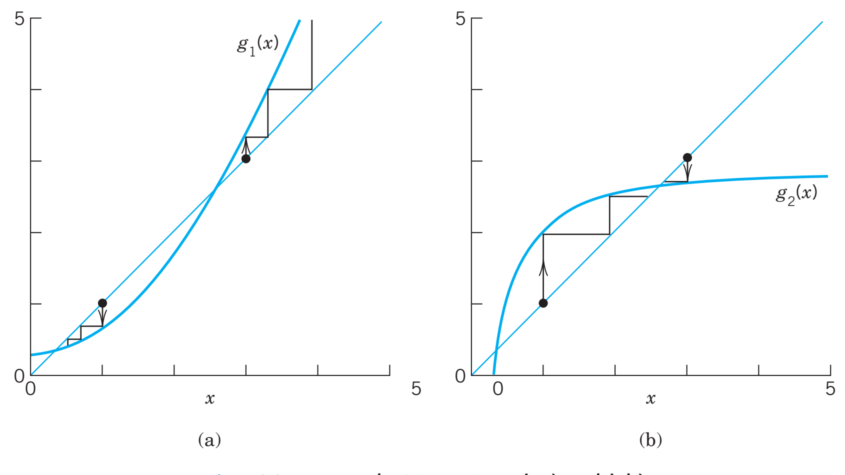


Fig. 426. Example 1, iterations (4a) and (4b)

Our figures show the following. In the lower part of Fig. 426a the slope of $g_1(x)$ is less than the slope of $y = x$, which is 1, thus $|g_1'(x)| < 1$, and we seem to have convergence. In the upper part, $g_1(x)$ is steeper ($g_1'(x) > 1$) and we have divergence. In Fig. 426b the slope of $g_2(x)$ is less near the intersection point ($x = 2.618$, fixed point of g_2, solution of $f(x) = 0$), and both sequences seem to converge. From all this we conclude that convergence seems to depend on the fact that, in a neighborhood of a solution, the curve of $g(x)$ is less steep than the straight line $y = x$, and we shall now see that this condition $|g'(x)| < 1$ (= slope of $y = x$) is sufficient for convergence. ■

An iteration process defined by (3) is called **convergent** for an x_0 if the corresponding sequence $x_0, x_1, \cdots$ is convergent.

A sufficient condition for convergence is given in the following theorem, which has various practical applications.

THEOREM 1

Convergence of Fixed-Point Iteration

Let $x = s$ be a solution of $x = g(x)$ and suppose that g has a continuous derivative in some interval J containing s. Then, if $|g'(x)| \leqq K < 1$ in J, the iteration process defined by (3) *converges for any x_0 in J. The limit of the sequence $\{x_n\}$ is s.*

PROOF By the mean value theorem of differential calculus there is a t between x and s such that

$$g(x) - g(s) = g'(t)(x - s) \qquad (x \text{ in } J).$$

Since $g(s) = s$ and $x_1 = g(x_0), x_2 = g(x_1), \cdots$, we obtain from this and the condition on $|g'(x)|$ in the theorem

$$|x_n - s| = |g(x_{n-1}) - g(s)| = |g'(t)||x_{n-1} - s| \leqq K|x_{n-1} - s|.$$

Applying this inequality n times, for $n, n - 1, \cdots, 1$ gives

$$|x_n - s| \leqq K|x_{n-1} - s| \leqq K^2|x_{n-2} - s| \leqq \cdots \leqq K^n|x_0 - s|.$$

Since $K < 1$, we have $K^n \to 0$; hence $|x_n - s| \to 0$ as $n \to \infty$. ■

We mention that a function g satisfying the condition in Theorem 1 is called a **contraction** because $|g(x) - g(v)| \leqq K|x - v|$, where $K < 1$. Furthermore, K gives information on the speed of convergence. For instance, if $K = 0.5$, then the accuracy increases by at least 2 digits in only 7 steps because $0.5^7 < 0.01$.

EXAMPLE 2

An Iteration Process. Illustration of Theorem 1

Find a solution of $f(x) = x^3 + x - 1 = 0$ by iteration.

Solution. A sketch shows that a solution lies near $x = 1$. **(a)** We may write the equation as $(x^2 + 1)x = 1$ or

$$x = g_1(x) = \frac{1}{1 + x^2}, \qquad \text{so that} \qquad x_{n+1} = \frac{1}{1 + x_n^2}. \qquad \text{Also} \qquad |g_1'(x)| = \frac{2|x|}{(1 + x^2)^2} < 1$$

for any x because $4x^2/(1 + x^2)^4 = 4x^2/(1 + 4x^2 + \cdots) < 1$, so that by Theorem 1 we have convergence for any x_0. Choosing $x_0 = 1$, we obtain (Fig. 427)

$$x_1 = 0.500, \quad x_2 = 0.800, \quad x_3 = 0.610, \quad x_4 = 0.729, \quad x_5 = 0.653, \quad x_6 = 0.701, \cdots.$$

The solution exact to 6D is $s = 0.682328$.

(b) The given equation may also be written

$$x = g_2(x) = 1 - x^3. \qquad \text{Then} \qquad |g_2'(x)| = 3x^2$$

and this is greater than 1 near the solution, so that we cannot apply Theorem 1 and assert convergence. Try $x_0 = 1, x_0 = 0.5, x_0 = 2$ and see what happens.

The example shows that the transformation of a given $f(x) = 0$ into the form $x = g(x)$ with g satisfying $|g'(x) \leqq K < 1$ may need some experimentation. ■

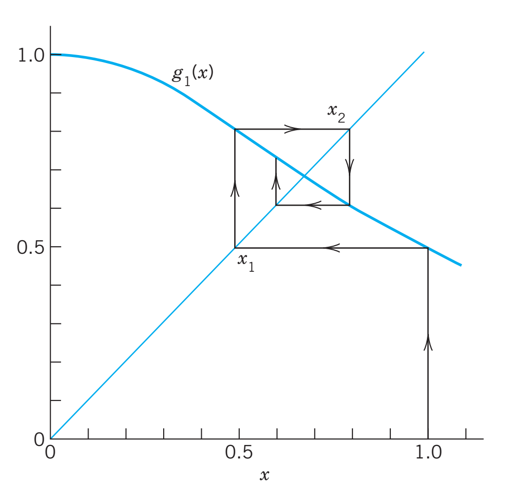


Fig. 427. Iteration in Example 2

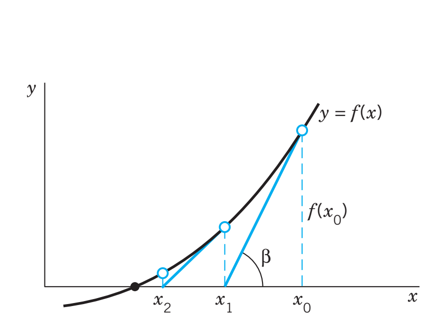


Fig. 428. Newton's method

Newton's Method for Solving Equations $f(x) = 0$

Newton's method, also known as **Newton–Raphson's method**,[1] is another iteration method for solving equations $f(x) = 0$, where f is assumed to have a continuous derivative f'. The method is commonly used because of its simplicity and great speed.

The underlying idea is that we approximate the graph of f by suitable tangents. Using an approximate value x_0 obtained from the graph of f, we let x_1 be the point of intersection of the x-axis and the tangent to the curve of f at x_0 (see Fig. 428). Then

$$\tan \beta = f'(x_0) = \frac{f(x_0)}{x_0 - x_1}, \qquad \text{hence} \qquad x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}.$$

In the second step we compute $x_2 = x_1 - f(x_1)/f'(x_1)$, in the third step x_3 from x_2 again by the same formula, and so on. We thus have the algorithm shown in Table 19.1. Formula (5) in this algorithm can also be obtained if we algebraically solve Taylor's formula

$$(5^*) \qquad f(x_{n+1}) \approx f(x_n) + (x_{n+1} - x_n)f'(x_n) = 0.$$

---

[1] JOSEPH RAPHSON (1648–1715), English mathematician who published a method similar to Newton's method. For historical details, see Ref. [GenRef2], p. 203, listed in App. 1.

Table 19.1 Newton's Method for Solving Equations $f(x) = 0$

ALGORITHM NEWTON $(f, f', x_0, \epsilon, N)$

This algorithm computes a solution of $f(x) = 0$ given an initial approximation x_0 (starting value of the iteration). Here the function $f(x)$ is continuous and has a continuous derivative $f'(x)$.

INPUT: f, f', initial approximation x_0, tolerance $\epsilon > 0$, maximum number of iterations N.

OUTPUT: Approximate solution x_n $(n \leqq N)$ or message of failure.

For $n = 0, 1, 2, \cdots, N - 1$ do:

1 Compute $f'(x_n)$.

2 If $f'(x_n) = 0$ then OUTPUT "Failure." Stop.

[*Procedure completed unsuccessfully*]

3 Else compute

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{5}$$

4 If $|x_{n+1} - x_n| \leqq \epsilon |x_{n+1}|$ then OUTPUT x_{n+1}. Stop.

[*Procedure completed successfully*]

End

5 OUTPUT "Failure". Stop.

[*Procedure completed unsuccessfully after N iterations*]

End NEWTON

If it happens that $f'(x_n) = 0$ for some n (see line 2 of the algorithm), then try another starting value x_0. Line 3 is the heart of Newton's method.

The inequality in line 4 is a **termination criterion**. If the sequence of the x_n converges and the criterion holds, we have reached the desired accuracy and stop. Note that this is just a form of the relative error test. It ensures that the result has the desired number of significant digits. If $|x_{n+1}| = 0$, the condition is satisfied if and only if $x_{n+1} = x_n = 0$, otherwise $|x_{n+1} - x_n|$ must be sufficiently small. The factor $|x_{n+1}|$ is needed in the case of zeros of very small (or very large) absolute value because of the high density (or of the scarcity) of machine numbers for those x.

WARNING! The criterion by itself does not imply convergence. *Example.* The harmonic series diverges, although its partial sums $x_n = \sum_{k=1}^{n} 1/k$ satisfy the criterion because $\lim (x_{n+1} - x_n) = \lim (1/(n + 1)) = 0$.

Line 5 gives another termination criterion and is needed because Newton's method may diverge or, due to a poor choice of x_0, may not reach the desired accuracy by a reasonable number of iterations. Then we may try another x_0. If $f(x) = 0$ has more than one solution, different choices of x_0 may produce different solutions. Also, an iterative sequence may sometimes converge to a solution different from the expected one.

EXAMPLE 3 **Square Root**

Set up a Newton iteration for computing the square root x of a given positive number c and apply it to $c = 2$.

Solution. We have $x = \sqrt{c}$, hence $f(x) = x^2 - c = 0$, $f'(x) = 2x$, and (5) takes the form

$$x_{n+1} = x_n - \frac{x_n^2 - c}{2x_n} = \frac{1}{2}\left(x_n + \frac{c}{x_n}\right).$$

For $c = 2$, choosing $x_0 = 1$, we obtain

$$x_1 = 1.500000, \qquad x_2 = 1.416667, \qquad x_3 = 1.414216, \qquad x_4 = 1.414214, \cdots.$$

x_4 is exact to 6D. ■

EXAMPLE 4 **Iteration for a Transcendental Equation**

Find the positive solution of $2 \sin x = x$.

Solution. Setting $f(x) = x - 2 \sin x$, we have $f'(x) = 1 - 2 \cos x$, and (5) gives

$$x_{n+1} = x_n - \frac{x_n - 2 \sin x_n}{1 - 2 \cos x_n} = \frac{2(\sin x_n - x_n \cos x_n)}{1 - 2 \cos x_n} = \frac{N_n}{D_n}.$$

n	x_n	N_n	D_n	x_{n+1}
0	2.00000	3.48318	1.83229	1.90100
1	1.90100	3.12470	1.64847	1.89552
2	1.89552	3.10500	1.63809	1.89550
3	1.89550	3.10493	1.63806	1.89549

From the graph of f we conclude that the solution is near $x_0 = 2$. We compute: $x_4 = 1.89549$ is exact to 5D since the solution to 6D is 1.895494. ■

EXAMPLE 5 **Newton's Method Applied to an Algebraic Equation**

Apply Newton's method to the equation $f(x) = x^3 + x - 1 = 0$.

Solution. From (5) we have

$$x_{n+1} = x_n - \frac{x_n^3 + x_n - 1}{3x_n^2 + 1} = \frac{2x_n^3 + 1}{3x_n^2 + 1}.$$

Starting from $x_0 = 1$, we obtain

$$x_1 = 0.750000, \quad x_2 = 0.686047, \quad x_3 = 0.682340, \quad x_4 = 0.682328, \cdots$$

where x_4 has the error $-1 \cdot 10^{-6}$. A comparison with Example 2 shows that the present convergence is much more rapid. This may motivate the concept of the *order of an iteration process,* to be discussed next. ■

Order of an Iteration Method. Speed of Convergence

The quality of an iteration method may be characterized by the speed of convergence, as follows.

Let $x_{n+1} = g(x_n)$ define an iteration method, and let x_n approximate a solution s of $x = g(x)$. Then $x_n = s - \epsilon_n$, where ϵ_n is the error of x_n. Suppose that g is differentiable a number of times, so that the Taylor formula gives

$$\begin{aligned} x_{n+1} = g(x_n) &= g(s) + g'(s)(x_n - s) + \tfrac{1}{2} g''(s)(x_n - s)^2 + \cdots \\ &= g(s) - g'(s)\epsilon_n + \tfrac{1}{2} g''(s)\epsilon_n^2 + \cdots . \end{aligned} \tag{6}$$

The exponent of ϵ_n in the first nonvanishing term after $g(s)$ is called the **order** of the iteration process defined by g. The order measures the speed of convergence.

To see this, subtract $g(s) = s$ on both sides of (6). Then on the left you get $x_{n+1} - s = -\epsilon_{n+1}$, where ϵ_{n+1} is the error of x_{n+1}. And on the right the remaining expression equals approximately its first nonzero term because $|\epsilon_n|$ is small in the case of convergence. Thus

$$\begin{aligned} &\text{(a)} \quad \epsilon_{n+1} \approx +g'(s)\epsilon_n && \text{in the case of first order,} \\ &\text{(b)} \quad \epsilon_{n+1} \approx -\tfrac{1}{2} g''(s)\epsilon_n^2 && \text{in the case of second order,} \qquad \text{etc.} \end{aligned} \tag{7}$$

Thus if $\epsilon_n = 10^{-k}$ in some step, then for second order, $\epsilon_{n+1} = c \cdot (10^{-k})^2 = c \cdot 10^{-2k}$, so that the number of significant digits is about doubled in each step.

Convergence of Newton's Method

In Newton's method, $g(x) = x - f(x)/f'(x)$. By differentiation,

$$\begin{aligned} g'(x) &= 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} \\ &= \frac{f(x)f''(x)}{f'(x)^2} . \end{aligned} \tag{8}$$

Since $f(s) = 0$, this shows that also $g'(s) = 0$. Hence Newton's method is at least of second order. If we differentiate again and set $x = s$, we find that

$$g''(s) = \frac{f''(s)}{f'(s)} \tag{8*}$$

which will not be zero in general. This proves

THEOREM 2

Second-Order Convergence of Newton's Method

If $f(x)$ is three times differentiable and f' and f'' are not zero at a solution s of $f(x) = 0$, then for x_0 sufficiently close to s, Newton's method is of second order.

Comments. For Newton's method, (7b) becomes, by (8*),

$$\epsilon_{n+1} \approx -\frac{f''(s)}{2f'(s)}\epsilon_n^2. \tag{9}$$

For the rapid convergence of the method indicated in Theorem 2 it is important that s be a *simple* zero of $f(x)$ (thus $f'(s) \neq 0$) and that x_0 be close to s, because in Taylor's formula we took only the linear term [see (5*)], assuming the quadratic term to be negligibly small. (With a bad x_0 the method may even diverge!)

EXAMPLE 6 Prior Error Estimate of the Number of Newton Iteration Steps

Use $x_0 = 2$ and $x_1 = 1.901$ in Example 4 for estimating how many iteration steps we need to produce the solution to 5D-accuracy. This is an **a priori estimate** or **prior estimate** because we can compute it after only one iteration, prior to further iterations.

Solution. We have $f(x) = x - 2\sin x = 0$. Differentiation gives

$$\frac{f''(s)}{2f'(s)} \approx \frac{f''(x_1)}{2f'(x_1)} = \frac{2\sin x_1}{2(1 - 2\cos x_1)} \approx 0.57.$$

Hence (9) gives

$$|\epsilon_{n+1}| \approx 0.57\epsilon_n^2 \approx 0.57(0.57\epsilon_{n-1}^2)^2 = 0.57^3\epsilon_{n-1}^4 \approx \cdots \approx 0.57^M\epsilon_0^{M+1} \leqq 5\cdot 10^{-6}$$

where $M = 2^n + 2^{n-1} + \cdots + 2 + 1 = 2^{n+1} - 1$. We show below that $\epsilon_0 \approx -0.11$. Consequently, our condition becomes

$$0.57^M 0.11^{M+1} \leqq 5\cdot 10^{-6}.$$

Hence $n = 2$ is the smallest possible n, according to this crude estimate, in good agreement with Example 4.

$\epsilon_0 \approx -0.11$ is obtained from $\epsilon_1 - \epsilon_0 = (\epsilon_1 - s) - (\epsilon_0 - s) = -x_1 + x_0 \approx 0.10$, hence $\epsilon_1 = \epsilon_0 + 0.10 \approx -0.57\epsilon_0^2$ or $0.57\epsilon_0^2 + \epsilon_0 + 0.10 \approx 0$, which gives $\epsilon_0 \approx -0.11$. ■

Difficulties in Newton's Method. Difficulties may arise if $|f'(x)|$ is very small near a solution s of $f(x) = 0$. For instance, let s be a zero of $f(x)$ of second or higher order. Then Newton's method converges only linearly, as is shown by an application of l'Hopital's rule to (8). Geometrically, small $|f'(x)|$ means that the tangent of $f(x)$ near s almost coincides with the x-axis (so that double precision may be needed to get $f(x)$ and $f'(x)$ accurately enough). Then for values $x = \tilde{s}$ far away from s we can still have small function values

$$R(\tilde{s}) = f(\tilde{s}).$$

In this case we call the equation $f(x) = 0$ **ill-conditioned**. $R(\tilde{s})$ is called the **residual** of $f(x) = 0$ at $\tilde{s}$. Thus a small residual guarantees a small error of $\tilde{s}$ only if the equation is ***not*** ill-conditioned.

EXAMPLE 7 An Ill-Conditioned Equation

$f(x) = x^5 + 10^{-4}x = 0$ is ill-conditioned, $x = 0$ is a solution. $f'(0) = 10^{-4}$ is small. At $\tilde{s} = 0.1$ the residual $f(0.1) = 2\cdot 10^{-5}$ is small, but the error -0.1 is larger in absolute value by a factor 5000. Invent a more drastic example of your own. ■

Secant Method for Solving $f(x) = 0$

Newton's method is very powerful but has the disadvantage that the derivative f' may sometimes be a far more difficult expression than f itself and its evaluation therefore

computationally expensive. This situation suggests the idea of replacing the derivative with the difference quotient

$$f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}} .$$

Then instead of (5) we have the formula of the popular secant method

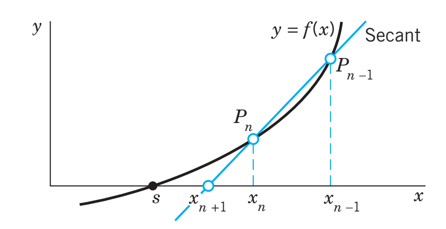


Fig. 429. Secant method

$$x_{n+1} = x_n - f(x_n)\frac{x_n - x_{n-1}}{f(x_n) - f(x_{n-1})} . \quad \textbf{(10)}$$

Geometrically, we intersect the x-axis at x_{n+1} with the secant of $f(x)$ passing through P_{n-1} and P_n in Fig. 429. We need two starting values x_0 and x_1. Evaluation of derivatives is now avoided. It can be shown that convergence is **superlinear** (that is, more rapid than linear, $|\epsilon_{n+1}| \approx \text{const} \cdot |\epsilon_n|^{1.62}$; see [E5] in App. 1), almost quadratic like Newton's method. The algorithm is similar to that of Newton's method, as the student may show.

CAUTION! It is ***not*** good to write (10) as

$$x_{n+1} = \frac{x_{n-1}f(x_n) - x_n f(x_{n-1})}{f(x_n) - f(x_{n-1})} ,$$

because this may lead to loss of significant digits if x_n and x_{n-1} are about equal. (Can you see this from the formula?)

EXAMPLE 8 Secant Method

Find the positive solution of $f(x) = x - 2 \sin x = 0$ by the secant method, starting from $x_0 = 2$, $x_1 = 1.9$.

Solution. Here, (10) is

$$x_{n+1} = x_n - \frac{(x_n - 2\sin x_n)(x_n - x_{n-1})}{x_n - x_{n-1} + 2(\sin x_{n-1} - \sin x_n)} = x_n - \frac{N_n}{D_n} .$$

Numeric values are:

n	x_{n-1}	x_n	N_n	D_n	$x_{n+1} - x_n$
1	2.000000	1.900000	−0.000740	−0.174005	−0.004253
2	1.900000	1.895747	−0.000002	−0.006986	−0.000252
3	1.895747	1.895494	0		0

$x_3 = 1.895494$ is exact to 6D. See Example 4. ■

Summary of Methods. The methods for computing solutions s of $f(x) = 0$ with given continuous (or differentiable) $f(x)$ start with an initial approximation x_0 of s and generate a sequence $x_1, x_2, \cdots$ by **iteration**. **Fixed-point methods** solve $f(x) = 0$ written as $x = g(x)$, so that s is a *fixed point* of g, that is, $s = g(s)$. For $g(x) = x - f(x)/f'(x)$ this is **Newton's method**, which, for good x_0 and simple zeros, converges quadratically (and for multiple zeros linearly). From Newton's method the **secant method** follows by replacing $f'(x)$ by a difference quotient. The ***bisection method*** and the ***method of false position*** in Problem Set 19.2 always converge, but often slowly.

PROBLEM SET 19.2

1–13 FIXED-POINT ITERATION

Solve by fixed-point iteration and answer related questions where indicated. Show details.

1. **Monotone sequence.** Why is the sequence in Example 1 monotone? Why not in Example 2?
2. Do the iterations (b) in Example 2. Sketch a figure similar to Fig. 427. Explain what happens.
3. $f = x - 0.5 \cos x = 0, \quad x_0 = 1$. Sketch a figure.
4. $f = x - \operatorname{cosec} x$ the zero near $x = 1$.
5. Sketch $f(x) = x^3 - 5.00x^2 + 1.01x + 1.88$, showing roots near ± 1 and 5. Write $x = g(x) = (5.00x^2 - 1.01x + 1.88)/x^2$. Find a root by starting from $x_0 = 5, 4, 1, -1$. Explain the (perhaps unexpected) results.
6. Find a form $x = g(x)$ of $f(x) = 0$ in Prob. 5 that yields convergence to the root near $x = 1$.
7. Find the smallest positive solution of $\sin x = e^{-x}$.
8. Solve $x^4 - x - 0.12 = 0$ by starting from $x_0 = 1$.
9. Find the negative solution of $x^4 - x - 0.12 = 0$.
10. **Elasticity.** Solve $x \cosh x = 1$. (Similar equations appear in vibrations of beams; see Problem Set 12.3.)
11. **Drumhead. Bessel functions.** A partial sum of the Maclaurin series of $J_0(x)$ (Sec. 5.5) is $f(x) = 1 - \frac{1}{4}x^2 + \frac{1}{64}x^4 - \frac{1}{2304}x^6$. Conclude from a sketch that $f(x) = 0$ near $x = 2$. Write $f(x) = 0$ as $x = g(x)$ (by dividing $f(x)$ by $\frac{1}{4}x$ and taking the resulting x-term to the other side). Find the zero. (See Sec. 12.10 for the importance of these zeros.)
12. **CAS EXPERIMENT. Convergence.** Let $f(x) = x^3 + 2x^2 - 3x - 4 = 0$. Write this as $x = g(x)$, for g choosing (1) $(x^3 - f)^{1/3}$, (2) $(x^2 - \frac{1}{2}f)^{1/2}$, (3) $x + \frac{1}{3}f$, (4) $(x^3 - f)/x^2$, (5) $(2x^2 - f)/(2x)$, and (6) $x - f/f'$ and in each case $x_0 = 1.5$. Find out about convergence and divergence and the number of steps to reach 6S-values of a root.
13. **Existence of fixed point.** Prove that if g is continuous in a closed interval I and its range lies in I, then the equation $x = g(x)$ has at least one solution in I. Illustrate that it may have more than one solution in I.

14–23 NEWTON'S METHOD

Apply Newton's method (6S-accuracy). First sketch the function(s) to see what is going on.

14. **Cube root.** Design a Newton iteration. Compute $\sqrt[3]{7}$, $x_0 = 2$.
15. $f = 2x - \cos x, \quad x_0 = 1$. Compare with Prob. 3.
16. What happens in Prob. 15 for any other x_0?
17. **Dependence on x_0.** Solve Prob. 5 by Newton's method with $x_0 = 5, 4, 1, -3$. Explain the result.
18. **Legendre polynomials.** Find the largest root of the Legendre polynomial $P_5(x)$ given by $P_5(x) = \frac{1}{8}(63x^5 - 70x^3 + 15x)$ (Sec. 5.3) (to be needed in ***Gauss integration*** in Sec. 19.5) **(a)** by Newton's method, **(b)** from a quadratic equation.
19. **Associated Legendre functions.** Find the smallest positive zero of $P_4^2 = (1 - x^2)P_4'' = \frac{15}{2}(-7x^4 + 8x^2 - 1)$ (Sec. 5.3) **(a)** by Newton's method, **(b)** exactly, by solving a quadratic equation.
20. $x + \ln x = 2, \quad x_0 = 2$
21. $f = x^3 - 5x + 3 = 0, \quad x_0 = 2, \quad 0, -2$
22. **Heating, cooling.** At what time x (4S-accuracy only) will the processes governed by $f_1(x) = 100(1 - e^{-0.2x})$ and $f_2(x) = 40e^{-0.01x}$ reach the same temperature? Also find the latter.
23. **Vibrating beam.** Find the solution of $\cos x \cosh x = 1$ near $x = \frac{3}{2}\pi$. (This determines a frequency of a vibrating beam; see Problem Set 12.3.)
24. **Method of False Position (Regula falsi).** Figure 430 shows the idea. We assume that f is continuous. We compute the x-intercept c_0 of the line through $(a_0, f(a_0))$, $(b_0, f(b_0))$. If $f(c_0) = 0$, we are done. If $f(a_0)f(c_0) < 0$ (as in Fig. 430), we set $a_1 = a_0$, $b_1 = c_0$ and repeat to get c_1, etc. If $f(a_0)f(c_0) > 0$, then $f(c_0)f(b_0) < 0$ and we set $a_1 = c_0$, $b_1 = b_0$, etc.

    (a) Algorithm. Show that

    $$c_0 = \frac{a_0 f(b_0) - b_0 f(a_0)}{f(b_0) - f(a_0)}$$

    and write an algorithm for the method.

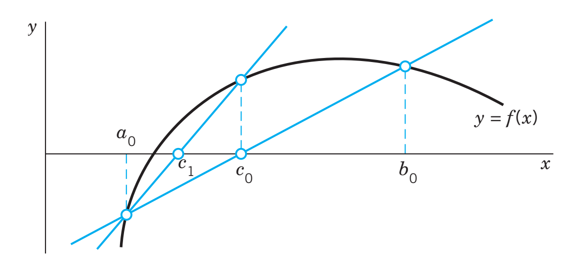


Fig. 430. Method of false position

(b) Solve $x^4 = 2$, $\cos x = \sqrt{x}$, and $x + \ln x = 2$, with $a = 1$, $b = 2$.

25. TEAM PROJECT. Bisection Method. This simple but slowly convergent method for finding a solution of $f(x) = 0$ with continuous f is based on the **intermediate value theorem**, which states that if a continuous function f has opposite signs at some $x = a$ and $x = b\,(> a)$, that is, either $f(a) < 0, f(b) > 0$ or $f(a) > 0, f(b) < 0$, then f must be 0 somewhere on $[a, b]$. The solution is found by repeated bisection of the interval and in each iteration picking that half which also satisfies that sign condition.

(a) Algorithm. Write an algorithm for the method.

(b) Comparison. Solve $x = \cos x$ by Newton's method and by bisection. Compare.

(c) Solve $e^{-x} = \ln x$ and $e^x + x^4 + x = 2$ by bisection.

26–29 SECANT METHOD

Solve, using x_0 and x_1 as indicated:

26. $e^{-x} - \tan x = 0, \quad x_0 = 1, \quad x_1 = 0.7$

27. Prob. 21, $x_0 = 1.0, \quad x_1 = 2.0$

28. $x = \cos x, \quad x_0 = 0.5, \quad x_1 = 1$

29. $\sin x = \cot x, \quad x_0 = 1, \quad x_1 = 0.5$

30. WRITING PROJECT. Solution of Equations. Compare the methods in this section and problem set, discussing advantages and disadvantages in terms of examples of your own. No proofs, just motivations and ideas.

19.3 Interpolation

We are given the values of a function $f(x)$ at different points $x_0, x_1, \cdots, x_n$. We want to find approximate values of the function $f(x)$ for "new" x's that lie between these points for which the function values are given. This process is called **interpolation**. The student should pay close attention to this section as interpolation forms the underlying foundation for both Secs. 19.4 and 19.5. Indeed, interpolation allows us to develop formulas for numeric integration and differentiation as shown in Sec. 19.5.

Continuing our discussion, we write these given values of a function f in the form

$$f_0 = f(x_0), \qquad f_1 = f(x_1), \qquad \cdots, \qquad f_n = f(x_n)$$

or as ordered pairs

$$(x_0, f_0), \qquad (x_1, f_1), \qquad \cdots, \quad (x_n, f_n).$$

Where do these given function values come from? They may come from a "mathematical" function, such as a logarithm or a Bessel function. More frequently, they may be measured or automatically recorded values of an "empirical" function, such as air resistance of a car or an airplane at different speeds. Other examples of functions that are "empirical" are the yield of a chemical process at different temperatures or the size of the U.S. population as it appears from censuses taken at 10-year intervals.

A standard idea in interpolation now is to find a polynomial $p_n(x)$ of degree n (or less) that assumes the given values; thus

$$\textbf{(1)} \qquad p_n(x_0) = f_0, \qquad p_n(x_1) = f_1, \qquad \cdots, \qquad p_n(x_n) = f_n.$$

We call this p_n an **interpolation polynomial** and $x_0, \cdots, x_n$ the **nodes**. And if $f(x)$ is a mathematical function, we call p_n an **approximation** of f (or a **polynomial approximation**, because there are other kinds of approximations, as we shall see later). We use p_n to get (approximate) values of f for x's between x_0 and x_n ("**interpolation**") or sometimes outside this interval $x_0 \leqq x \leqq x_n$ ("**extrapolation**").

Motivation. Polynomials are convenient to work with because we can readily differentiate and integrate them, again obtaining polynomials. Moreover, they approximate *continuous* functions with any desired accuracy. That is, for any continuous $f(x)$ on an interval J: $a \leqq x \leqq b$ and error bound $\beta > 0$, there is a polynomial $p_n(x)$ (of sufficiently high degree n) such that

$$|f(x) - p_n(x)| < \beta \quad \text{for all } x \text{ on } J.$$

This is the famous **Weierstrass approximation theorem** (for a proof see Ref. [GenRef7], App. 1).

Existence and Uniqueness. Note that the interpolation polynomial p_n satisfying (1) for given data exists and we shall give formulas for it below. Furthermore, p_n is unique: Indeed, if another polynomial q_n also satisfies $q_n(x_0) = f_0, \cdots, q_n(x_n) = f_n$, then $p_n(x) - q_n(x) = 0$ at $x_0, \cdots, x_n$, but a polynomial $p_n - q_n$ of degree n (or less) with $n + 1$ roots must be identically zero, as we know from algebra; thus $p_n(x) = q_n(x)$ for all x, which means uniqueness. ■

How Do We Find p_n? We shall explain several standard methods that give us p_n. By the uniqueness proof above, we know that, for given data, the different methods *must* give us the same polynomial. However, the polynomials may be expressed in different forms suitable for different purposes.

Lagrange Interpolation

Given $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with arbitrarily spaced x_j, Lagrange had the idea of multiplying each f_j by a polynomial that is 1 at x_j and 0 at the other n nodes and then taking the sum of these $n + 1$ polynomials. Clearly, this gives the unique interpolation polynomial of degree n or less. Beginning with the simplest case, let us see how this works.

Linear interpolation is interpolation by the straight line through $(x_0, f_0), (x_1, f_1)$; see Fig. 431. Thus the linear Lagrange polynomial p_1 is a sum $p_1 = L_0 f_0 + L_1 f_1$ with L_0 the linear polynomial that is 1 at x_0 and 0 at x_1; similarly, L_1 is 0 at x_0 and 1 at x_1. Obviously,

$$L_0(x) = \frac{x - x_1}{x_0 - x_1}, \qquad L_1(x) = \frac{x - x_0}{x_1 - x_0}.$$

This gives the linear Lagrange polynomial

$$p_1(x) = L_0(x) f_0 + L_1(x) f_1 = \frac{x - x_1}{x_0 - x_1} \cdot f_0 + \frac{x - x_0}{x_1 - x_0} \cdot f_1. \tag{2}$$

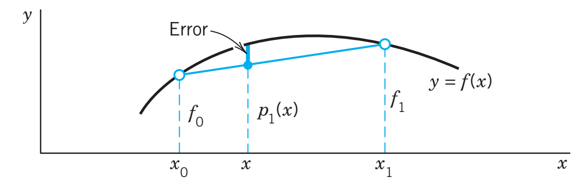


Fig. 431. Linear Interpolation

EXAMPLE 1 **Linear Lagrange Interpolation**

Compute a 4D-value of ln 9.2 from ln 9.0 = 2.1972, ln 9.5 = 2.2513 by linear Lagrange interpolation and determine the error, using ln 9.2 = 2.2192 (4D).

Solution. $x_0 = 9.0, x_1 = 9.5, f_0 = \ln 9.0, f_1 = \ln 9.5$. Ln (2) we need

$$L_0(x) = \frac{x - 9.5}{-0.5} = -2.0(x - 9.5), \qquad L_0(9.2) = -2.0(-0.3) = 0.6$$

$$L_1(x) = \frac{x - 9.0}{0.5} = 2.0(x - 9.0), \qquad L_1(9.2) = 2 \cdot 0.2 = 0.4$$

(see Fig. 432) and obtain the answer

$$\ln 9.2 \approx p_1(9.2) = L_0(9.2)\,f_0 + L_1(9.2)\,f_1 = 0.6 \cdot 2.1972 + 0.4 \cdot 2.2513 = 2.2188.$$

The error is $\epsilon = a - \tilde{a} = 2.2192 - 2.2188 = 0.0004$. Hence linear interpolation is not sufficient here to get 4D accuracy; it would suffice for 3D accuracy. ■

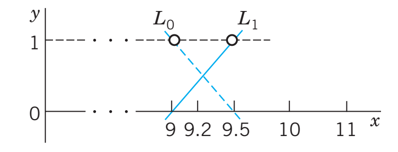


Fig. 432. L_0 and L_1 in Example 1

Quadratic interpolation is interpolation of given $(x_0, f_0), (x_1, f_1), (x_2, f_2)$ by a second-degree polynomial $p_2(x)$, which by Lagrange's idea is

$$p_2(x) = L_0(x)f_0 + L_1(x)f_1 + L_2(x)f_2 \tag{3a}$$

with $L_0(x_0) = 1, L_1(x_1) = 1, L_2(x_2) = 1$, and $L_0(x_1) = L_0(x_2) = 0$, etc. We claim that

$$\begin{aligned} L_0(x) &= \frac{l_0(x)}{l_0(x_0)} = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} \\ L_1(x) &= \frac{l_1(x)}{l_1(x_1)} = \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} \\ L_2(x) &= \frac{l_2(x)}{l_2(x_2)} = \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)}. \end{aligned} \tag{3b}$$

How did we get this? Well, the numerator makes $L_k(x_j) = 0$ if $j \neq k$. And the denominator makes $L_k(x_k) = 1$ because it equals the numerator at $x = x_k$.

EXAMPLE 2 **Quadratic Lagrange Interpolation**

Compute ln 9.2 by (3) from the data in Example 1 and the additional third value ln 11.0 = 2.3979.

Solution. In (3),

$$L_0(x) = \frac{(x - 9.5)(x - 11.0)}{(9.0 - 9.5)(9.0 - 11.0)} = x^2 - 20.5x + 104.5, \qquad L_0(9.2) = 0.5400,$$

$$L_1(x) = \frac{(x - 9.0)(x - 11.0)}{(9.5 - 9.0)(9.5 - 11.0)} = -\frac{1}{0.75}(x^2 - 20x + 99), \qquad L_1(9.2) = 0.4800,$$

$$L_2(x) = \frac{(x - 9.0)(x - 9.5)}{(11.0 - 9.0)(11.0 - 9.5)} = \frac{1}{3}(x^2 - 18.5x + 85.5), \qquad L_2(9.2) = -0.0200,$$

(see Fig. 433), so that (3a) gives, exact to 4D,

$$\ln 9.2 \approx p_2(9.2) = 0.5400 \cdot 2.1972 + 0.4800 \cdot 2.2513 - 0.0200 \cdot 2.3979 = 2.2192.$$ ■

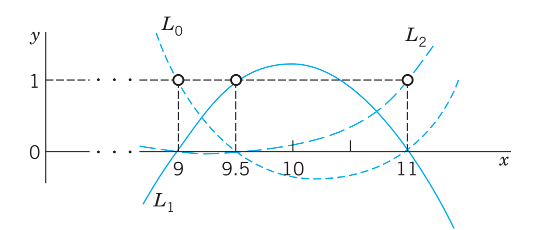


Fig. 433. L_0, L_1, L_2 in Example 2

General Lagrange Interpolation Polynomial. For general n we obtain

(4a)
$$f(x) \approx p_n(x) = \sum_{k=0}^{n} L_k(x) f_k = \sum_{k=0}^{n} \frac{l_k(x)}{l_k(x_k)} f_k$$

where $L_k(x_k) = 1$ and L_k is 0 at the other nodes, and the L_k are independent of the function f to be interpolated. We get (4a) if we take

(4b)
$$\begin{aligned} l_0(x) &= (x - x_1)(x - x_2) \cdots (x - x_n), \\ l_k(x) &= (x - x_0) \cdots (x - x_{k-1})(x - x_{k+1}) \cdots (x - x_n), \qquad 0 < k < n, \\ l_n(x) &= (x - x_0)(x - x_1) \cdots (x - x_{n-1}). \end{aligned}$$

We can easily see that $p_n(x_k) = f_k$. Indeed, inspection of (4b) shows that $l_k(x_j) = 0$ if $j \neq k$, so that for $x = x_k$, the sum in (4a) reduces to the single term $(l_k(x_k)/l_k(x_k)) f_k = f_k$.

Error Estimate. If f is itself a polynomial of degree n (or less), it must coincide with p_n because the $n + 1$ data $(x_0, f_0), \cdots, (x_n, f_n)$ determine a polynomial uniquely, so the error is zero. Now the special f has its $(n + 1)$st derivative identically zero. This makes it plausible that for a *general* f its $(n + 1)$st derivative $f^{(n+1)}$ should measure the error

$$\epsilon_n(x) = f(x) - p_n(x).$$

It can be shown that this is true if $f^{(n+1)}$ exists and is continuous. Then, with a suitable t between x_0 and x_n (or between x_0, x_n, and x if we extrapolate),

(5)
$$\epsilon_n(x) = f(x) - p_n(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(t)}{(n + 1)!}.$$

Thus $|\epsilon_n(x)|$ is 0 at the nodes and small near them, because of continuity. The product $(x - x_0) \cdots (x - x_n)$ is large for x away from the nodes. This makes extrapolation risky. And interpolation at an x will be best if we choose nodes on both sides of that x. Also, we get error bounds by taking the smallest and the largest value of $f^{(n+1)}(t)$ in (5) on the interval $x_0 \leqq t \leqq x_n$ (or on the interval also containing x if we *extra*polate).

Most importantly, since p_n is unique, as we have shown, we have

THEOREM 1

Error of Interpolation

Formula (5) *gives the error for* ***any*** *polynomial interpolation method if* $f(x)$ *has a continuous* $(n + 1)$*st derivative.*

Practical error estimate. If the derivative in (5) is difficult or impossible to obtain, apply the Error Principle (Sec. 19.1), that is, take another node and the Lagrange polynomial $p_{n+1}(x)$ and regard $p_{n+1}(x) - p_n(x)$ as a (crude) error estimate for $p_n(x)$.

EXAMPLE 3 Error Estimate (5) of Linear Interpolation. Damage by Roundoff. Error Principle

Estimate the error in Example 1 first by (5) directly and then by the Error Principle (Sec. 19.1).

Solution. **(A)** ***Estimation by*** **(5).** We have $n = 1$, $f(t) = \ln t$, $f'(t) = 1/t$, $f''(t) = -1/t^2$. Hence

$$\epsilon_1(x) = (x - 9.0)(x - 9.5)\frac{(-1)}{2t^2}, \qquad \text{thus} \qquad \epsilon_1(9.2) = \frac{0.03}{t^2}.$$

$t = 0.9$ gives the maximum $0.03/9^2 = 0.00037$ and $t = 9.5$ gives the minimum $0.03/9.5^2 = 0.00033$, so that we get $0.00033 \leqq \epsilon_1(9.2) \leqq 0.00037$, or better, 0.00038 because $0.3/81 = 0.003703 \cdots$.

But the error 0.0004 in Example 1 disagrees, and we can learn something! Repetition of the computation there with 5D instead of 4D gives

$$\ln 9.2 \approx p_1(9.2) = 0.6 \cdot 2.19722 + 0.4 \cdot 2.25129 = 2.21885$$

with an actual error $\epsilon = 2.21920 - 2.21885 = 0.00035$, which lies nicely near the middle between our two error bounds.

This shows that the discrepancy (0.0004 vs. 0.00035) was caused by rounding, which is not taken into account in (5).

(B) ***Estimation by the Error Principle.*** We calculate $p_1(9.2) = 2.21885$ as before and then $p_2(9.2)$ as in Example 2 but with 5D, obtaining

$$p_2(9.2) = 0.54 \cdot 2.19722 + 0.48 \cdot 2.25129 - 0.02 \cdot 2.39790 = 2.21916.$$

The difference $p_2(9.2) - p_1(9.2) = 0.00031$ is the approximate error of $p_1(9.2)$ that we wanted to obtain; this is an approximation of the actual error 0.00035 given above. ■

Newton's Divided Difference Interpolation

For given data $(x_0, f_0), \cdots, (x_n, f_n)$ the interpolation polynomial $p_n(x)$ satisfying (1) is unique, as we have shown. But for different purposes we may use $p_n(x)$ in different forms. **Lagrange's form** just discussed is useful for deriving formulas in numeric differentiation (approximation formulas for derivatives) and integration (Sec. 19.5).

Practically more important are Newton's forms of $p_n(x)$, which we shall also use for solving ODEs (in Sec. 21.2). They involve fewer arithmetic operations than Lagrange's form. Moreover, it often happens that we have to increase the degree n to reach a required accuracy. Then in Newton's forms we can use all the previous work and just add another term, a possibility without counterpart for Lagrange's form. This also simplifies the application of the Error Principle (used in Example 3 for Lagrange). The details of these ideas are as follows.

Let $p_{n-1}(x)$ be the $(n - 1)$st Newton polynomial (whose form we shall determine); thus $p_{n-1}(x_0) = f_0$, $p_{n-1}(x_1) = f_1, \cdots, p_{n-1}(x_{n-1}) = f_{n-1}$. Furthermore, let us write the nth Newton polynomial as

$$p_n(x) = p_{n-1}(x) + g_n(x); \tag{6}$$

hence

$$g_n(x) = p_n(x) - p_{n-1}(x). \tag{6'}$$

Here $g_n(x)$ is to be determined so that $p_n(x_0) = f_0, p_n(x_1) = f_1, \cdots, p_n(x_n) = f_n$.

Since p_n and p_{n-1} agree at $x_0, \cdots, x_{n-1}$, we see that g_n is zero there. Also, g_n will generally be a polynomial of nth degree because so is p_n, whereas p_{n-1} can be of degree $n - 1$ at most. Hence g_n must be of the form

$$g_n(x) = a_n(x - x_0)(x - x_1) \cdots (x - x_{n-1}). \tag{6''}$$

We determine the constant a_n. For this we set $x = x_n$ and solve (6″) algebraically for a_n. Replacing $g_n(x_n)$ according to (6′) and using $p_n(x_n) = f_n$, we see that this gives

$$a_n = \frac{f_n - p_{n-1}(x_n)}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})}. \tag{7}$$

We write a_k instead of a_n and show that a_k equals the **kth divided difference**, recursively denoted and defined as follows:

$$a_1 = f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0}$$

$$a_2 = f[x_0, x_1, x_2] = \frac{f[x_1, x_2] - f[x_0, x_1]}{x_2 - x_0}$$

and in general

$$a_k = f[x_0, \cdots, x_k] = \frac{f[x_1, \cdots, x_k] - f[x_0, \cdots, x_{k-1}]}{x_k - x_0}. \tag{8}$$

If $n = 1$, then $p_{n-1}(x_n) = p_0(x_1) = f_0$ because $p_0(x)$ is constant and equal to f_0, the value of $f(x)$ at x_0. Hence (7) gives

$$a_1 = \frac{f_1 - p_0(x_1)}{x_1 - x_0} = \frac{f_1 - f_0}{x_1 - x_0} = f[x_0, x_1],$$

and (6) and (6″) give the Newton interpolation polynomial of the first degree

$$p_1(x) = f_0 + (x - x_0)f[x_0, x_1].$$

If $n = 2$, then this p_1 and (7) give

$$a_2 = \frac{f_2 - p_1(x_2)}{(x_2 - x_0)(x_2 - x_1)} = \frac{f_2 - f_0 - (x_2 - x_0)f[x_0, x_1]}{(x_2 - x_0)(x_2 - x_1)} = f[x_0, x_1, x_2]$$

where the last equality follows by straightforward calculation and comparison with the definition of the right side. (Verify it; be patient.) From (6) and (6″) we thus obtain the second Newton polynomial

$$p_2(x) = f_0 + (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2].$$

For $n = k$, formula (6) gives

$$p_k(x) = p_{k-1}(x) + (x - x_0)(x - x_1) \cdots (x - x_{k-1})f[x_0, \cdots, x_k]. \tag{9}$$

With $p_0(x) = f_0$ by repeated application with $k = 1, \cdots, n$ this finally gives **Newton's divided difference interpolation formula**

$$\begin{aligned} f(x) \approx f_0 &+ (x - x_0)f[x_0, x_1] + (x - x_0)(x - x_1)f[x_0, x_1, x_2] \\ &+ \cdots + (x - x_0)(x - x_1) \cdots (x - x_{n-1})f[x_0, \cdots, x_n]. \end{aligned} \tag{10}$$

An algorithm is shown in Table 19.2. The first do-loop computes the divided differences and the second the desired value $p_n(\hat{x})$.

Example 4 shows how to arrange differences near the values from which they are obtained; the latter always stand a half-line above and a half-line below in the preceding column. Such an arrangement is called a (divided) **difference table**.

Table 19.2 Newton's Divided Difference Interpolation

ALGORITHM INTERPOL ($x_0, \cdots, x_n$; $f_0, \cdots, f_n$; $\hat{x}$)

This algorithm computes an approximation $p_n(\hat{x})$ of $f(\hat{x})$ at $\hat{x}$.

INPUT: Data $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$; $\hat{x}$

OUTPUT: Approximation $p_n(\hat{x})$ of $f(\hat{x})$

Set $f[x_j] = f_j \quad (j = 0, \cdots, n)$.

For $m = 1, \cdots, n - 1)$ do:

For $j = 0, \cdots, n - m$ do:

$$f[x_j, \cdots, x_{j+m}] = \frac{f[x_{j+1}, \cdots, x_{j+m}] - f[x_j, \cdots, x_{j+m-1}]}{x_{j+m} - x_j}$$

End

End

Set $p_0(x) = f_0$.

For $k = 1, \cdots, n$ do:

$$p_k(\hat{x}) = p_{k-1}(\hat{x}) + (\hat{x} - x_0) \cdots (\hat{x} - x_{k-1})f[x_0, \cdots, x_k]$$

End

OUTPUT $p_n(\hat{x})$

End INTERPOL

EXAMPLE 4 Newton's Divided Difference Interpolation Formula

Compute $f(9.2)$ from the values shown in the first two columns of the following table.

x_j	$f_j = f(x_j)$	$f[x_j, x_{j+1}]$	$f[x_j, x_{j+1}, x_{j+2}]$	$f[x_j, \cdots, x_{j+3}]$
8.0	(2.079442)			
		(0.117783)		
9.0	2.197225		(−0.006433)	
		0.108134		(0.000411)
9.5	2.251292		−0.005200	
		0.097735		
11.0	2.397895			

Solution. We compute the divided differences as shown. Sample computation:

$$(0.097735 - 0.108134)/(11 - 9) = -0.005200.$$

The values we need in (10) are circled. We have

$$f(x) \approx p_3(x) = 2.079442 + 0.117783(x - 8.0) - 0.006433(x - 8.0)(x - 9.0)$$
$$+ 0.000411(x - 8.0)(x - 9.0)(x - 9.5).$$

At $x = 9.2$,

$$f(9.2) \approx 2.079442 + 0.141340 - 0.001544 - 0.000030 = 2.219208.$$

The value exact to 6D is $f(9.2) = \ln 9.2 = 2.219203$. Note that we can nicely see how the accuracy increases from term to term:

$$p_1(9.2) = 2.220782, \quad p_2(9.2) = 2.219238, \quad p_3(9.2) = 2.219208.$$ ■

Equal Spacing: Newton's Forward Difference Formula

Newton's formula (10) is valid for *arbitrarily spaced* nodes as they may occur in practice in experiments or observations. However, in many applications the x_j's are *regularly spaced*—for instance, in measurements taken at regular intervals of time. Then, denoting the distance by h, we can write

(11) $$x_0, \quad x_1 = x_0 + h, \quad x_2 = x_0 + 2h, \quad \cdots, \quad x_n = x_0 + nh.$$

We show how (8) and (10) now simplify considerably!

To get started, let us define the *first forward difference* of f at x_j by

$$\Delta f_j = f_{j+1} - f_j,$$

the *second forward difference* of f at x_j by

$$\Delta^2 f_j = \Delta f_{j+1} - \Delta f_j,$$

and, continuing in this way, the ***k*th forward difference** of f at x_j by

(12) $$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 1, 2, \cdots).$$

Examples and an explanation of the name "forward" follow on the next page. What is the point of this? We show that if we have regular spacing (11), then

(13)
$$f[x_0, \cdots, x_k] = \frac{1}{k!h^k} \Delta^k f_0.$$

PROOF We prove (13) by induction. It is true for $k = 1$ because $x_1 = x_0 + h$, so that

$$f[x_0, x_1] = \frac{f_1 - f_0}{x_1 - x_0} = \frac{1}{h}(f_1 - f_0) = \frac{1}{1!h} \Delta f_0.$$

Assuming (13) to be true for all forward differences of order k, we show that (13) holds for $k + 1$. We use (8) with $k + 1$ instead of k; then we use $(k + 1)h = x_{k+1} - x_0$, resulting from (11), and finally (12) with $j = 0$, that is, $\Delta^{k+1} f_0 = \Delta^k f_1 - \Delta^k f_0$. This gives

$$\begin{aligned} f[x_0, \cdots, x_{k+1}] &= \frac{f[x_1, \cdots, x_{k+1}] - f[x_0, \cdots, x_k]}{(k + 1)h} \\ &= \frac{1}{(k + 1)h}\left[\frac{1}{k!h^k} \Delta^k f_1 - \frac{1}{k!h^k} \Delta^k f_0\right] \\ &= \frac{1}{(k + 1)!h^{k+1}} \Delta^{k+1} f_0 \end{aligned}$$

which is (13) with $k + 1$ instead of k. Formula (13) is proved. ■

In (10) we finally set $x = x_0 + rh$. Then $x - x_0 = rh$, $x - x_1 = (r - 1)h$ since $x_1 - x_0 = h$, and so on. With this and (13), formula (10) becomes **Newton's** (or *Gregory*[2]*–Newton's*) **forward difference interpolation formula**

(14)
$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r}{s} \Delta^s f_0 \qquad (x = x_0 + rh, \quad r = (x - x_0)/h) \\ &= f_0 + r\Delta f_0 + \frac{r(r - 1)}{2!} \Delta^2 f_0 + \cdots + \frac{r(r - 1)\cdots(r - n + 1)}{n!} \Delta^n f_0 \end{aligned}$$

where the **binomial coefficients** in the first line are defined by

(15)
$$\binom{r}{0} = 1, \quad \binom{r}{s} = \frac{r(r - 1)(r - 2)\cdots(r - s + 1)}{s!} \qquad (s > 0, \text{ integer})$$

and $s! = 1 \cdot 2 \cdots s$.

Error. From (5) we get, with $x - x_0 = rh$, $x - x_1 = (r - 1)h$, etc.,

(16)
$$\epsilon_n(x) = f(x) - p_n(x) = \frac{h^{n+1}}{(n + 1)!} r(r - 1) \cdots (r - n) f^{(n+1)}(t)$$

with t as characterized in (5).

[2] JAMES GREGORY (1638–1675), Scots mathematician, professor at St. Andrews and Edinburgh. Δ in (14) and ∇^2 (on p. 818) have nothing to do with the Laplacian.

Formula (16) is an exact formula for the error, but it involves the unknown t. In Example 5 (below) we show how to use (16) for obtaining an error estimate and an interval in which the true value of $f(x)$ must lie.

Comments on Accuracy. (A) The order of magnitude of the error $\epsilon_n(x)$ is about equal to that of the next difference not used in $p_n(x)$.

(B) One should choose $x_0, \cdots, x_n$ such that the x at which one interpolates is as well centered between $x_0, \cdots, x_n$ as possible.

The reason for (A) is that in (16),

$$f^{n+1}(t) \approx \frac{\Delta^{n+1} f(t)}{h^{n+1}}, \qquad \frac{|r(r-1)\cdots(r-n)|}{1 \cdot 2 \cdots (n+1)} \leqq 1 \quad \text{if} \quad |r| \leqq 1$$

(and actually for any r as long as we do not *extrapolate*). The reason for (B) is that $|r(r-1)\cdots(r-n)|$ becomes smallest for that choice.

EXAMPLE 5 Newton's Forward Difference Formula. Error Estimation

Compute cosh 0.56 from (14) and the four values in the following table and estimate the error.

j	x_j	$f_j = \cosh x_j$	Δf_j	$\Delta^2 f_j$	$\Delta^3 f_j$
0	0.5	1.127626			
			0.057839		
1	0.6	1.185465		0.011865	
			0.069704		0.000697
2	0.7	1.255169		0.012562	
			0.082266		
3	0.8	1.337435			

Solution. We compute the forward differences as shown in the table. The values we need are circled. In (14) we have $r = (0.56 - 0.50)/0.1 = 0.6$, so that (14) gives

$$\begin{aligned}\cosh 0.56 &\approx 1.127626 + 0.6 \cdot 0.057839 + \frac{0.6(-0.4)}{2} \cdot 0.011865 + \frac{0.6(-0.4)(-1.4)}{6} \cdot 0.000697 \\ &= 1.127626 + 0.034703 - 0.001424 + 0.000039 \\ &= 1.160944.\end{aligned}$$

Error estimate. From (16), since the fourth derivative is $\cosh^{(4)} t = \cosh t$,

$$\begin{aligned}\epsilon_3(0.56) &= \frac{0.1^4}{4!} \cdot 0.6(-0.4)(-1.4)(-2.4) \cosh t \\ &= A \cosh t,\end{aligned}$$

where $A = -0.00000336$ and $0.5 \leqq t \leqq 0.8$. We do not know t, but we get an inequality by taking the largest and smallest $\cosh t$ in that interval:

$$A \cosh 0.8 \leqq \epsilon_3(0.62) \leqq A \cosh 0.5.$$

Since

$$f(x) = p_3(x) + \epsilon_3(x),$$

this gives

$$p_3(0.56) + A \cosh 0.8 \leqq \cosh 0.56 \leqq p_3(0.56) + A \cosh 0.5.$$

Numeric values are

$$1.160939 \leqq \cosh 0.56 \leqq 1.160941.$$

The exact 6D-value is $\cosh 0.56 = 1.160941$. It lies within these bounds. Such bounds are not always so tight. Also, we did not consider roundoff errors, which will depend on the number of operations. ■

This example also explains the name "*forward* difference formula": we see that the differences in the formula slope forward in the difference table.

Equal Spacing: Newton's Backward Difference Formula

Instead of forward-sloping differences we may also employ backward-sloping differences. The difference table remains the same as before (same numbers, in the same positions), except for a very harmless change of the running subscript j (which we explain in Example 6, below). Nevertheless, purely for reasons of convenience it is standard to introduce a second name and notation for differences as follows. We define the *first backward difference* of f at x_j by

$$\nabla f_j = f_j - f_{j-1},$$

the *second backward difference* of f at x_j by

$$\nabla^2 f_j = \nabla f_j - \nabla f_{j-1},$$

and, continuing in this way, the ***k*th backward difference** of f at x_j by

$$\nabla^k f_j = \nabla^{k-1} f_j - \nabla^{k-1} f_{j-1} \qquad (k = 1, 2, \cdots). \tag{17}$$

A formula similar to (14) but involving backward differences is **Newton's** (or *Gregory–Newton's*) **backward difference interpolation formula**

$$\begin{aligned} f(x) \approx p_n(x) &= \sum_{s=0}^{n} \binom{r+s-1}{s} \nabla^s f_0 \qquad (x = x_0 + rh,\ r = (x - x_0)/h) \\ &= f_0 + r\nabla f_0 + \frac{r(r+1)}{2!} \nabla^2 f_0 + \cdots + \frac{r(r+1)\cdots(r+n-1)}{n!} \nabla^n f_0. \end{aligned} \tag{18}$$

EXAMPLE 6 Newton's Forward and Backward Interpolations

Compute a 7D-value of the Bessel function $J_0(x)$ for $x = 1.72$ from the four values in the following table, using (a) Newton's forward formula (14), (b) Newton's backward formula (18).

j_{for}	j_{back}	x_j	$J_0(x_j)$	1st Diff.	2nd Diff.	3rd Diff.
0	−3	1.7	0.3979849			
				−0.0579985		
1	−2	1.8	0.3399864		−0.0001693	
				−0.0581678		0.0004093
2	−1	1.9	0.2818186		0.0002400	
				−0.0579278		
3	0	2.0	0.2238908			

Solution. The computation of the differences is the same in both cases. Only their notation differs.

(a) Forward. In (14) we have $r = (1.72 - 1.70)/0.1 = 0.2$, and j goes from 0 to 3 (see first column). In each column we need the first given number, and (14) thus gives

$$J_0(1.72) \approx 0.3979849 + 0.2(-0.0579985) + \frac{0.2(-0.8)}{2}(-0.0001693) + \frac{0.2(-0.8)(-1.8)}{6} \cdot 0.0004093$$
$$= 0.3979849 - 0.0115997 + 0.0000135 + 0.0000196 = 0.3864183,$$

which is exact to 6D, the exact 7D-value being 0.3864185.

(b) Backward. For (18) we use j shown in the second column, and in each column the last number. Since $r = (1.72 - 2.00)/0.1 = -2.8$, we thus get from (18)

$$J_0(1.72) \approx 0.2238908 - 2.8(-0.0579278) + \frac{-2.8(-1.8)}{2} \cdot 0.0002400 + \frac{-2.8(-1.8)(-0.8)}{6} \cdot 0.0004093$$
$$= 0.2238908 + 0.1621978 + 0.0006048 - 0.0002750$$
$$= 0.3864184.$$

There is a third notation for differences, called the **central difference notation**. It is used in numerics for ODEs and certain interpolation formulas. See Ref. [E5] listed in App. 1.

PROBLEM SET 19.3

1. **Linear interpolation.** Calculate $p_1(x)$ in Example 1 and from it ln 9.3.
2. **Error estimate.** Estimate the error in Prob. 1 by (5).
3. **Quadratic interpolation. Gamma function.** Calculate the Lagrange polynomial $p_2(x)$ for the values $\Gamma(1.00) = 1.0000$, $\Gamma(1.02) = 0.9888$, $\Gamma(1.04) = 0.9784$ of the gamma function [(24) in App. A3.1] and from it approximations of $\Gamma(1.01)$ and $\Gamma(1.03)$.
4. **Error estimate for quadratic interpolation.** Estimate the error for $p_2(9.2)$ in Example 2 from (5).
5. **Linear and quadratic interpolation.** Find $e^{-0.25}$ and $e^{-0.75}$ by linear interpolation of e^{-x} with $x_0 = 0$, $x_1 = 0.5$ and $x_0 = 0.5$, $x_1 = 1$, respectively. Then find $p_2(x)$ by quadratic interpolation of e^{-x} with $x_0 = 0$, $x_1 = 0.5$, $x_2 = 1$ and from it $e^{-0.25}$ and $e^{-0.75}$. Compare the errors. Use 4S-values of e^{-x}.
6. **Interpolation and extrapolation.** Calculate $p_2(x)$ in Example 2. Compute from it approximations of ln 9.4, ln 10, ln 10.5, ln 11.5, and ln 12. Compute the errors by using exact 5S-values and comment.
7. **Interpolation and extrapolation.** Find the quadratic polynomial that agrees with $\sin x$ at $x = 0, \pi/4, \pi/2$ and use it for the interpolation and extrapolation of $\sin x$ at $x = -\pi/8, \pi/8, 3\pi/8, 5\pi/8$. Compute the errors.
8. **Extrapolation.** Does a sketch of the product of the $(x - x_j)$ in (5) for the data in Example 2 indicate that extrapolation is likely to involve larger errors than interpolation does?
9. **Error function** (35) in App. A3.1. Calculate the Lagrange polynomial $p_2(x)$ for the 5S-values $f(0.25) = 0.27633$, $f(0.5) = 0.52050$, $f(1.0) = 0.84270$ and from $p_2(x)$ an approximation of $f(0.75)$ $(= 0.71116)$.

10. Error bound. Derive an error bound in Prob. 9 from (5).

11. Cubic Lagrange interpolation. Bessel function J_0. Calculate and graph L_0, L_1, L_2, L_3 with $x_0 = 0$, $x_1 = 1, x_2 = 2, x_3 = 3$ on common axes. Find $p_3(x)$ for the data (0, 1), (1, 0.765198), (2, 0.223891), (3, −0.260052) [values of the Bessel function $J_0(x)$]. Find p_3 for $x = 0.5, 1.5, 2.5$ and compare with the 6S-exact values 0.938470, 0.511828, −0.048384.

12. Newton's forward formula (14). Sine integral. Using (14), find $f(1.25)$ by linear, quadratic, and cubic interpolation of the data (values of (40) in App. A31); 6S-value $\text{Si}(1.25) = 1.14645)$ $f(1.0) = 0.94608$, $f(1.5) = 1.32468$, $f(2.0) = 1.60541$, $f(2.5) = 1.77852$, and compute the errors. For the linear interpolation use $f(1.0)$ and $f(1.5)$, for the quadratic $f(1.0)$, $f(1.5)$, $f(2.0)$, etc.

13 Lower degree. Find the degree of the interpolation polynomial for the data (−4, 50), (−2, 18), (0, 2), (2, 2), (4, 18), using a difference table. Find the polynomial.

14. Newton's forward formula (14). Gamma function. Set up (14) for the data in Prob. 3 and compute $\Gamma(1.01)$, $\Gamma(1.03)$, $\Gamma(1.05)$.

15. Divided differences. Obtain p_2 in Example 2 from (10).

16. Divided differences. Error function. Compute $p_2(0.75)$ from the data in Prob. 9 and Newton's divided difference formula (10).

17. Backward difference formula (18). Use $p_2(x)$ in (18) and the values of erf x, $x = 0.2, 0.4, 0.6$ in Table A4 of App. 5, compute erf 0.3 and the error. (4S-exact erf 0.3 = 0.3286).

18. In Example 5 of the text, write down the difference table as needed for (18), then write (18) with general x and then with $x = 0.56$ to verify the answer in Example 5.

19. CAS EXPERIMENT. Adding Terms in Newton Formulas. Write a program for the forward formula (14). Experiment on the increase of accuracy by successively adding terms. As data use values of some function of your choice for which your CAS gives the values needed in determining errors.

20. TEAM PROJECT. Interpolation and Extrapolation.
(a) Lagrange practical error estimate (after Theorem 1). Apply this to $p_1(9.2)$ and $p_2(9.2)$ for the data $x_0 = 9.0, x_1 = 9.5, x_2 = 11.0, f_0 = \ln x_0, f_1 = \ln x_1$, $f_2 = \ln x_2$ (6S-values).

(b) Extrapolation. Given $(x_j, f(x_j)) = (0.2, 0.9980)$, (0.4, 0.9686), (0.6, 0.8443), (0.8, 0.5358), (1.0, 0). Find $f(0.7)$ from the quadratic interpolation polynomials based on (α) 0.6, 0.8, 1.0, (β) 0.4, 0.6, 0.8, (γ) 0.2, 0.4, 0.6. Compare the errors and comment. [Exact $f(x) = \cos(\frac{1}{2}\pi x^2)$, $f(0.7) = 0.7181$ (4S).]

(c) Graph the product of factors $(x - x_j)$ in the error formula (5) for $n = 2, \cdots, 10$ separately. What do these graphs show regarding accuracy of interpolation and extrapolation?

21. WRITING PROJECT. Comparison of interpolation methods. List 4–5 ideas that you feel are most important in this section. Arrange them in best logical order. Discuss them in a 2–3 page report.

19.4 Spline Interpolation

Given data (function values, points in the *xy*-plane) $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ can be interpolated by a polynomial $P_n(x)$ of degree n or less so that the curve of $P_n(x)$ passes through these $n + 1$ points (x_j, f_j); here $f_0 = f(x_0), \cdots, f_n = f(x_n)$, See Sec. 19.3.

Now if n is large, there may be trouble: $P_n(x)$ may tend to oscillate for x between the **nodes** $x_0, \cdots, x_n$. Hence we must be prepared for ***numeric instability*** (Sec. 19.1). Figure 434 shows a famous example by C. Runge[3] for which the maximum error even approaches ∞ as $n \to \infty$ (with the nodes kept equidistant and their number increased). Figure 435 illustrates the increase of the oscillation with n for some other function that is piecewise linear.

Those undesirable oscillations are avoided by the method of splines initiated by I. J. Schoenberg in 1946 (*Quarterly of Applied Mathematics* **4**, pp. 45–99, 112–141). This method is widely used in practice. It also laid the foundation for much of modern **CAD (computer-aided design)**. Its name is borrowed from a *draftman's spline*, which is an elastic rod bent to pass through given points and held in place by weights. The mathematical idea of the method is as follows:

[3]CARL RUNGE (1856–1927), German mathematician, also known for his work on ODEs (Sec. 21.1).

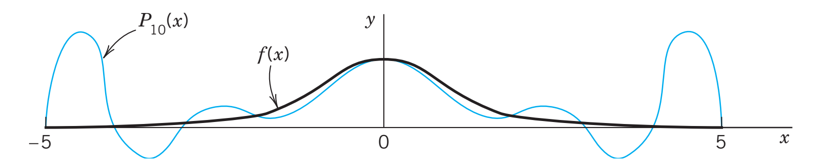


Fig. 434. Runge's example $f(x) = 1/(1 + x^2)$ and interpolating polynomial $P_{10}(x)$

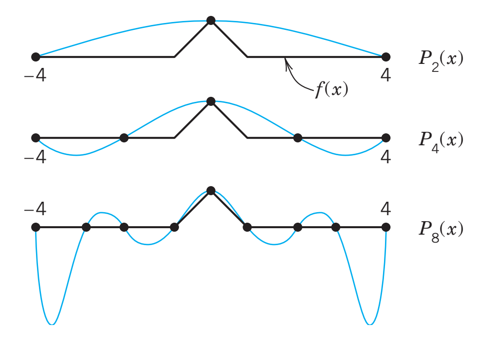


Fig. 435. Piecewise linear function $f(x)$ and interpolation polynomials of increasing degrees

Instead of using a single high-degree polynomial P_n over the entire interval $a \leqq x \leqq b$ in which the nodes lie, that is,

$$a = x_0 < x_1 < \cdots < x_n = b, \tag{1}$$

we use n low-degree, e.g., cubic, polynomials

$$q_0(x), \qquad q_1(x), \qquad \cdots, \qquad q_{n-1}(x),$$

one over each subinterval between adjacent nodes, hence q_0 from x_0 to x_1, then q_1 from x_1 to x_2, and so on. From this we compose an interpolation function $g(x)$, called a **spline**, by fitting these polynomials together into a single continuous curve passing through the data points, that is,

$$g(x_0) = f(x_0) = f_0, \qquad g(x_1) = f(x_1) = f_1, \qquad \cdots, \qquad g(x_n) = f(x_n) = f_n. \tag{2}$$

Note that $g(x) = q_0(x)$ when $x_0 \leqq x \leqq x_1$, then $g(x) = q_1(x)$ when $x_1 \leqq x \leqq x_2$, and so on, according to our construction of g.

Thus **spline interpolation** is *piecewise polynomial interpolation.*

The simplest q_j's would be linear polynomials. However, the curve of a piecewise linear continuous function has corners and would be of little interest in general—think of designing the body of a car or a ship.

We shall consider cubic splines because these are the most important ones in applications. By definition, a **cubic spline** $g(x)$ interpolating given data $(x_0, f_0), \cdots, (x_n, f_n)$ is a continuous function on the interval $a = x_0 \leqq x \leqq x_n = b$ that has continuous first and second derivatives and satisfies the interpolation condition (2); furthermore, between adjacent nodes, $g(x)$ is given by a polynomial $q_j(x)$ of degree 3 or less.

We claim that there is such a cubic spline. And if in addition to (2) we also require that

$$g'(x_0) = k_0, \qquad g'(x_n) = k_n \tag{3}$$

(given tangent directions of $g(x)$ at the two endpoints of the interval $a \leqq x \leqq b$), then we have a uniquely determined cubic spline. This is the content of the following existence and uniqueness theorem, whose proof will also suggest the actual determination of splines. (Condition (3) will be discussed after the proof.)

THEOREM 1

Existence and Uniqueness of Cubic Splines

Let $(x_0, f_0), (x_1, f_1), \cdots, (x_n, f_n)$ with given (arbitrarily spaced) x_j [see (1)] and given $f_j = f(x_j), j = 0, 1, \cdots, n$. Let k_0 and k_n be any given numbers. Then there is one and only one cubic spline $g(x)$ corresponding to (1) and satisfying (2) and (3).

PROOF By definition, on every subinterval I_j given by $x_j \leqq x \leqq x_{j+1}$, the spline $g(x)$ must agree with a polynomial $q_j(x)$ of degree not exceeding 3 such that

$$q_j(x_j) = f(x_j), \qquad q_j(x_{j+1}) = f(x_{j+1}) \qquad (j = 0, 1, \cdots, n-1). \tag{4}$$

For the derivatives we write

$$q_j'(x_j) = k_j, \qquad q_j'(x_{j+1}) = k_{j+1} \qquad (j = 0, 1, \cdots, n-1) \tag{5}$$

with k_0 and k_n given and $k_1, \cdots, k_{n-1}$ to be determined later. Equations (4) and (5) are four conditions for each $q_j(x)$. By direct calculation, using the notation

$$c_j = \frac{1}{h_j} = \frac{1}{x_{j+1} - x_j} \qquad (j = 0, 1, \cdots, n-1) \tag{6*}$$

we can verify that the unique cubic polynomial $q_j(x)$ $(j = 0, 1, \cdots, n-1)$ satisfying (4) and (5) is

$$\begin{aligned} q_j(x) = {} & f(x_j)c_j^2(x - x_{j+1})^2[1 + 2c_j(x - x_j)] \\ & + f(x_{j+1})c_j^2(x - x_j)^2[1 - 2c_j(x - x_{j+1})] \\ & + k_j c_j^2(x - x_j)(x - x_{j+1})^2 \\ & + k_{j+1}c_j^2(x - x_j)^2(x - x_{j+1}). \end{aligned} \tag{6}$$

Differentiating twice, we obtain

$$q_j''(x_j) = -6c_j^2 f(x_j) + 6c_j^2 f(x_{j+1}) - 4c_j k_j - 2c_j k_{j+1} \tag{7}$$

$$q_j''(x_{j+1}) = 6c_j^2 f(x_j) - 6c_j^2 f(x_{j+1}) + 2c_j k_j + 4c_j k_{j+1}. \tag{8}$$

By definition, $g(x)$ has continuous second derivatives. This gives the conditions

$$q_{j-1}''(x_j) = q_j''(x_j) \qquad (j = 1, \cdots, n-1).$$

If we use (8) with j replaced by $j - 1$, and (7), these $n - 1$ equations become

$$\textbf{(9)} \qquad c_{j-1}k_{j-1} + 2(c_{j-1} + c_j)k_j + c_jk_{j+1} = 3[c_{j-1}^2\nabla f_j + c_j^2\nabla f_{j+1}]$$

where $\nabla f_j = f(x_j) - f(x_{j-1})$ and $\nabla f_{j+1} = f(x_{j+1}) - f(x_j)$ and $j = 1, \cdots, n - 1$, as before. This linear system of $n - 1$ equations has a unique solution $k_1, \cdots, k_{n-1}$ since the coefficient matrix is strictly diagonally dominant (that is, in each row the (positive) diagonal entry is greater than the sum of the other (positive) entries). Hence the determinant of the matrix cannot be zero (as follows from Theorem 3 in Sec. 20.7), so that we may determine unique values $k_1, \cdots, k_{n-1}$ of the first derivative of $g(x)$ at the nodes. This proves the theorem. ■

Storage and Time Demands in solving (9) are modest, since the matrix of (9) is **sparse** (has few nonzero entries) and **tridiagonal** (may have nonzero entries only on the diagonal and on the two adjacent "parallels" above and below it). Pivoting (Sec. 7.3) is not necessary because of that dominance. This makes splines efficient in solving large problems with thousands of nodes or more. For some literature and some critical comments, see *American Mathematical Monthly* **105** (1998), 929–941.

Condition (3) includes the **clamped conditions**

$$(10) \qquad g'(x_0) = f'(x_0), \qquad g'(x_n) = f'(x_n),$$

in which the tangent directions $f'(x_0)$ and $f'(x_n)$ at the ends are given. Other conditions of practical interest are the **free** or **natural conditions**

$$(11) \qquad g''(x_0) = 0, \qquad g''(x_n) = 0$$

(geometrically: zero curvature at the ends, as for the draftman's spline), giving a **natural spline**. These names are motivated by Fig. 293 in Problem Set 12.3.

Determination of Splines. Let k_0 and k_n be given. Obtain $k_1, \cdots, k_{n-1}$ by solving the linear system (9). Recall that the spline $g(x)$ to be found consists of n cubic polynomials $q_0, \cdots, q_{n-1}$. We write these polynomials in the form

$$\textbf{(12)} \qquad q_j(x) = a_{j0} + a_{j1}(x - x_j) + a_{j2}(x - x_j)^2 + a_{j3}(x - x_j)^3$$

where $j = 0, \cdots, n - 1$. Using Taylor's formula, we obtain

$$\textbf{(13)} \qquad \begin{aligned} a_{j0} &= q_j(x_j) = f_j && \text{by (2),} \\ a_{j1} &= q_j'(x_j) = k_j && \text{by (5),} \\ a_{j2} &= \frac{1}{2}\, q_j''(x_j) = \frac{3}{h_j^2}\,(f_{j+1} - f_j) - \frac{1}{h_j}\,(k_{j+1} + 2k_j) && \text{by (7),} \\ a_{j3} &= \frac{1}{6}\, q_j'''(x_j) = \frac{2}{h_j^3}\,(f_j - f_{j+1}) + \frac{1}{h_j^2}\,(k_{j+1} + k_j) \end{aligned}$$

with a_{j3} obtained by calculating $q_j''(x_{j+1})$ from (12) and equating the result to (8), that is,

$$q_j''(x_{j+1}) = 2a_{j2} + 6a_{j3}h_j = \frac{6}{h_j^2}(f_j - f_{j+1}) + \frac{2}{h_j}(k_j + 2k_{j+1}),$$

and now subtracting from this $2a_{j2}$ as given in (13) and simplifying.

Note that for ***equidistant nodes*** of distance $h_j = h$ we can write $c_j = c = 1/h$ in (6*) and have from (9) simply

(14) $$k_{j-1} + 4k_j + k_{j+1} = \frac{3}{h}(f_{j+1} - f_{j-1}) \qquad (j = 1, \cdots, n-1).$$

EXAMPLE 1 Spline Interpolation. Equidistant Nodes

Interpolate $f(x) = x^4$ on the interval $-1 \leqq x \leqq 1$ by the cubic spline $g(x)$ corresponding to the nodes $x_0 = -1$, $x_1 = 0$, $x_2 = 1$ and satisfying the clamped conditions $g'(-1) = f'(-1)$, $g'(1) = f'(1)$.

Solution. In our standard notation the given data are $f_0 = f(-1) = 1, f_1 = f(0) = 0, f_2 = f(1) = 1$. We have $h = 1$ and $n = 2$, so that our spline consists of $n = 2$ polynomials

$$q_0(x) = a_{00} + a_{01}(x+1) + a_{02}(x+1)^2 + a_{03}(x+1)^3 \qquad (-1 \leqq x \leqq 0),$$
$$q_1(x) = a_{10} + a_{11}x + a_{12}x^2 + a_{13}x^3 \qquad (0 \leqq x \leqq 1).$$

We determine the k_j from (14) (equidistance!) and then the coefficients of the spline from (13). Since $n = 2$, the system (14) is a single equation (with $j = 1$ and $h = 1$)

$$k_0 + 4k_1 + k_2 = 3(f_2 - f_0).$$

Here $f_0 = f_2 = 1$ (the value of x^4 at the ends) and $k_0 = -4$, $k_2 = 4$, the values of the derivative $4x^3$ at the ends -1 and 1. Hence

$$-4 + 4k_1 + 4 = 3(1-1) = 0, \qquad k_1 = 0.$$

From (13) we can now obtain the coefficients of q_0, namely, $a_{00} = f_0 = 1$, $a_{01} = k_0 = -4$, and

$$a_{02} = \frac{3}{1^2}(f_1 - f_0) - \frac{1}{1}(k_1 + 2k_0) = 3(0-1) - (0-8) = 5$$

$$a_{03} = \frac{2}{1^3}(f_0 - f_1) + \frac{1}{1^2}(k_1 + k_0) = 2(1-0) + (0-4) = -2.$$

Similarly, for the coefficients of q_1 we obtain from (13) the values $a_{10} = f_1 = 0$, $a_{11} = k_1 = 0$, and

$$a_{12} = 3(f_2 - f_1) - (k_2 + 2k_1) = 3(1-0) - (4+0) = -1$$
$$a_{13} = 2(f_1 - f_2) + (k_2 + k_1) = 2(0-1) + (4+0) = 2.$$

This gives the polynomials of which the spline $g(x)$ consists, namely,

$$g(x) = \begin{cases} q_0(x) = 1 - 4(x+1) + 5(x+1)^2 - 2(x+1)^3 = -x^2 - 2x^3 & \text{if} \quad -1 \leqq x \leqq 0 \\ q_1(x) = -x^2 + 2x^3 & \text{if} \quad 0 \leqq x \leqq 1. \end{cases}$$

Figure 436 shows $f(x)$ and this spline. Do you see that we could have saved over half of our work by using symmetry? ■

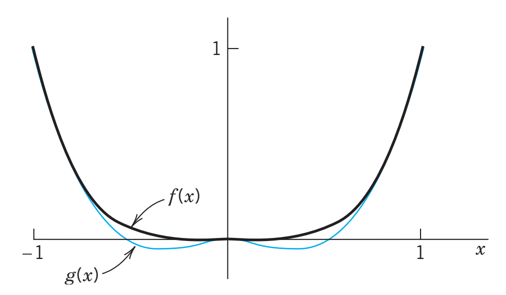


Fig. 436. Function $f(x) = x^4$ and cubic spline $g(x)$ in Example 1

EXAMPLE 2 Natural Spline. Arbitrarily Spaced Nodes

Find a spline approximation and a polynomial approximation for the curve of the cross section of the circular-shaped Shrine of the Book in Jerusalem shown in Fig. 437.

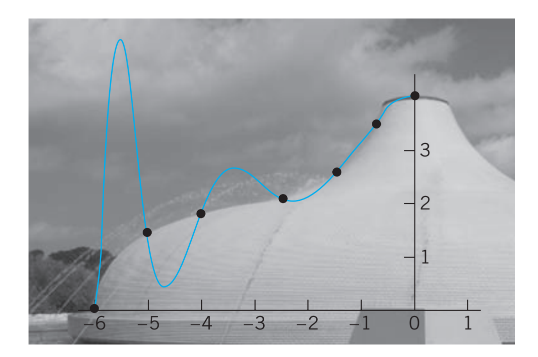


Fig. 437. Shrine of the Book in Jerusalem (Architects F. Kissler and A. M. Bartus)

Solution. Thirteen points, about equally distributed along the contour (not along the x-axis!), give these data:

x_j	−5.8	−5.0	−4.0	−2.5	−1.5	−0.8	0	0.8	1.5	2.5	4.0	5.0	5.8
f_j	0	1.5	1.8	2.2	2.7	3.5	3.9	3.5	2.7	2.2	1.8	1.5	0

The figure shows the corresponding interpolation polynomial of 12th degree, which is useless because of its oscillation. (Because of roundoff your software will also give you small error terms involving odd powers of x.) The polynomial is

$$P_{12}(x) = 3.9000 - 0.65083x^2 + 0.033858x^4 + 0.011041x^6 - 0.0014010x^8$$
$$+ 0.000055595x^{10} - 0.00000071867x^{12}.$$

The spline follows practically the contour of the roof, with a small error near the nodes −0.8 and 0.8. The spline is symmetric. Its six polynomials corresponding to positive x have the following coefficients of their representations (12). (Note well that (12) is in terms of powers of $x - x_j$, not x!)

j	x-interval	a_{j0}	a_{j1}	a_{j2}	a_{j3}
0	0.0...0.8	3.9	0.00	−0.61	−0.015
1	0.8...1.5	3.5	−1.01	−0.65	0.66
2	1.5...2.5	2.7	−0.95	0.73	−0.27
3	2.5...4.0	2.2	−0.32	−0.091	0.084
4	4.0...5.0	1.8	−0.027	0.29	−0.56
5	5.0...5.8	1.5	−1.13	−1.39	0.58

PROBLEM SET 19.4

1. WRITING PROJECT. Splines. In your own words, and using as few formulas as possible, write a short report on spline interpolation, its motivation, a comparison with polynomial interpolation, and its applications.

2–9 VERIFICATIONS. DERIVATIONS. COMPARISONS

2. Individual polynomial q_j. Show that $q_j(x)$ in (6) satisfies the interpolation condition (4) as well as the derivative condition (5).

3. Verify the differentiations that give (7) and (8) from (6).

4. System for derivatives. Derive the basic linear system (9) for $k_1, \cdots, k_{n-1}$ as indicated in the text.

5. Equidistant nodes. Derive (14) from (9).

6. Coefficients. Give the details of the derivation of a_{j2} and a_{j3} in (13).

7. Verify the computations in Example 1.

8. Comparison. Compare the spline g in Example 1 with the quadratic interpolation polynomial over the whole interval. Find the maximum deviations of g and p_2 from f. Comment.

9. Natural spline condition. Using the given coefficients, verify that the spline in Example 2 satisfies $g''(x) = 0$ at the ends.

10–16 DETERMINATION OF SPLINES

Find the cubic spline $g(x)$ for the given data with k_0 and k_n as given.

10. $f(-2) = f(-1) = f(1) = f(2) = 0, \quad f(0) = 1,$ $k_0 = k_4 = 0$

11. If we started from the piecewise linear function in Fig. 438, we would obtain $g(x)$ in Prob. 10 as the spline satisfying $g'(-2) = f'(-2) = 0, \quad g'(2) = f'(2) = 0.$ Find and sketch or graph the corresponding interpolation polynomial of 4th degree and compare it with the spline. Comment.

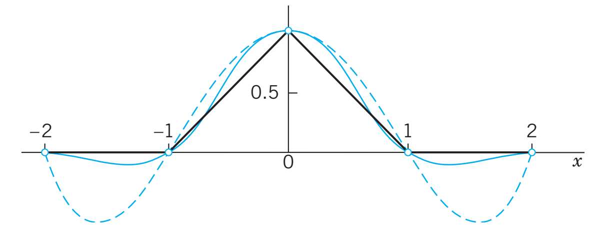


Fig. 438. Spline and interpolation polynomial in Probs. 10 and 11

12. $f_0 = f(0) = 1, \quad f_1 = f(2) = 9, \quad f_2 = f(4) = 41,$ $f_3 = f(6) = 41, \quad k_0 = 0, \quad k_3 = -12$

13. $f_0 = f(0) = 1, \quad f_1 = f(1) = 0, \quad f_2 = f(2) = -1,$ $f_3 = f(3) = 0, \quad k_0 = 0, \quad k_3 = -6$

14. $f_0 = f(0) = 2, \quad f_1 = f(1) = 3, \quad f_2 = f(2) = 8,$ $f_3 = f(3) = 12, \quad k_0 = k_3 = 0$

15. $f_0 = f(0) = 4, \quad f_1 = f(2) = 0, \quad f_2 = f(4) = 4,$ $f_3 = f(6) = 80, \quad k_0 = k_3 = 0$

16. $f_0 = f(0) = 2, \quad f_1 = f(2) = -2, \quad f_2 = f(4) = 2,$ $f_3 = f(6) = 78, \quad k_0 = k_3 = 0.$ Can you obtain the answer from that of Prob. 15?

17. If a cubic spline is three times continuously differentiable (that is, it has continuous first, second, and third derivatives), show that it must be a single polynomial.

18. CAS EXPERIMENT. Spline versus Polynomial. If your CAS gives natural splines, find the natural splines when x is integer from $-m$ to m, and $y(0) = 1$ and all other y equal to 0. Graph each such spline along with the interpolation polynomial p_{2m}. Do this for $m = 2$ to 10 (or more). What happens with increasing m?

19. Natural conditions. Explain the remark after (11).

20. TEAM PROJECT. Hermite Interpolation and Bezier Curves. In **Hermite interpolation** we are looking for a polynomial $p(x)$ (of degree $2n + 1$ or less) such that $p(x)$ and its derivative $p'(x)$ have given values at $n + 1$ nodes. (More generally, $p(x), p'(x), p''(x), \cdots$ may be required to have given values at the nodes.)

(a) Curves with given endpoints and tangents. Let C be a curve in the xy-plane parametrically represented by $\mathbf{r}(t) = [x(t), y(t)], 0 \leqq t \leqq 1$ (see Sec. 9.5). Show that for given initial and terminal points of a curve and given initial and terminal tangents, say,

$$
\begin{aligned}
A: \quad \mathbf{r}_0 &= [x(0), y(0)] \\
&= [x_0, y_0], \\
B: \quad \mathbf{r}_1 &= [x(1), y(1)] \\
&= [x_1, y_1] \\
\mathbf{v}_0 &= [x'(0), y'(0)] \\
&= [x_0', y_0'], \\
\mathbf{v}_1 &= [x'(1), y'(1)] \\
&= [x_1', y_1']
\end{aligned}
$$

we can find a curve C, namely,

$$
\begin{aligned}
\mathbf{r}(t) &= \mathbf{r}_0 + \mathbf{v}_0 t \\
&\quad + (3(\mathbf{r}_1 - \mathbf{r}_0) - (2\mathbf{v}_0 + \mathbf{v}_1))t^2 \\
&\quad + (2(\mathbf{r}_0 - \mathbf{r}_1) + \mathbf{v}_0 + \mathbf{v}_1)t^3;
\end{aligned}
\tag{15}
$$

in components,

$$x(t) = x_0 + x_0't + (3(x_1 - x_0) - (2x_0' + x_1'))t^2 + (2(x_0 - x_1) + x_0' + x_1')t^3$$

$$y(t) = y_0 + y_0't + (3(y_1 - y_0) - (2y_0' + y_1'))t^2 + (2(y_0 - y_1) + y_0' + y_1')t^3.$$

Note that this is a cubic Hermite interpolation polynomial, and $n = 1$ because we have two nodes (the endpoints of C). (This has nothing to do with the Hermite polynomials in Sec. 5.8.) The two points

$$G_A\colon\ \mathbf{g}_0 = \mathbf{r}_0 + \mathbf{v}_0 = [x_0 + x_0', y_0 + y_0']$$

and

$$G_B\colon\ \mathbf{g}_1 = \mathbf{r}_1 - \mathbf{v}_1 = [x_1 - x_1', y_1 - y_1']$$

are called **guidepoints** because the segments AG_A and BG_B specify the tangents graphically. A, B, G_A, G_B determine C, and C can be changed quickly by moving the points. A curve consisting of such Hermite interpolation polynomials is called a **Bezier curve**, after the French engineer P. Bezier of the Renault Automobile Company, who introduced them in the early 1960s in designing car bodies. Bezier curves (and surfaces) are used in computer-aided design (CAD) and computer-aided manufacturing (CAM). (For more details, see Ref. [E21] in App. 1.)

(b) Find and graph the Bezier curve and its guidepoints if A: [0, 0], B: [1, 0], $\mathbf{v}_0 = [\frac{1}{2}, \frac{1}{2}]$, $\mathbf{v}_1 = [-\frac{1}{2}, -\frac{1}{4}\sqrt{3}]$.

(c) Changing guidepoints changes C. Moving guidepoints farther away results in C "staying near the tangents for a longer time." Confirm this by changing $\mathbf{v}_0$ and $\mathbf{v}_1$ in (b) to $2\mathbf{v}_0$ and $2\mathbf{v}_1$ (see Fig. 439).

(d) Make experiments of your own. What happens if you change $\mathbf{v}_1$ in (b) to $-\mathbf{v}_1$. If you rotate the tangents? If you multiply $\mathbf{v}_0$ and $\mathbf{v}_1$ by positive factors less than 1?

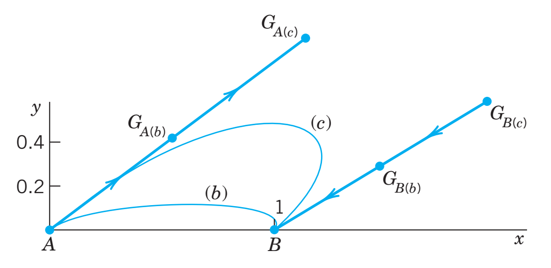


Fig. 439. Team Project 20(b) and (c): Bezier curves

19.5 Numeric Integration and Differentiation

In applications, the engineer often encounters integrals that are very difficult or even impossible to solve analytically. For example, the error function, the Fresnel integrals (see Probs. 16–25 on nonelementary integrals in this section), and others cannot be evaluated by the usual methods of calculus (see App. 3, (24)–(44) for such "difficult" integrals). We then need methods from numerical analysis to evaluate such integrals. We also need numerics when the integrand of the integral to be evaluated consists of an empirical function, where we are given some recorded values of that function. Methods that address these kinds of problems are called methods of numeric integration.

Numeric integration means the numeric evaluation of integrals

$$J = \int_a^b f(x)\,dx$$

where a and b are given and f is a function given analytically by a formula or empirically by a table of values. Geometrically, J is the area under the curve of f between a and b (Fig. 440), taken with a minus sign where f is negative.

We know that if f is such that we can find a differentiable function F whose derivative is f, then we can evaluate J directly, i.e., without resorting to numeric integration, by applying the familiar formula

$$J = \int_a^b f(x)\,dx = F(b) - F(a) \qquad [F'(x) = f(x)].$$

Your CAS (Mathematica, Maple, etc.) or tables of integrals may be helpful for this purpose.

Rectangular Rule. Trapezoidal Rule

Numeric integration methods are obtained by approximating the integrand f by functions that can easily be integrated.

The simplest formula, the **rectangular rule**, is obtained if we subdivide the interval of integration $a \leqq x \leqq b$ into n subintervals of equal length $h = (b - a)/n$ and in each subinterval approximate f by the constant $f(x_j^*)$, the value of f at the midpoint x_j^* of the jth subinterval (Fig. 441). Then f is approximated by a **step function** (piecewise constant function), the n rectangles in Fig. 441 have the areas $f(x_1^*)h, \cdots, f(x_n^*)h$, and the **rectangular rule** is

(1) $$J = \int_a^b f(x)\,dx \approx h[f(x_1^*) + f(x_2^*) + \cdots + f(x_n^*)] \qquad \left(h = \frac{b-a}{n}\right).$$

The **trapezoidal rule** is generally more accurate. We obtain it if we take the same subdivision as before and approximate f by a broken line of segments (chords) with endpoints $[a, f(a)], [x_1, f(x_1)], \cdots, [b, f(b)]$ on the curve of f (Fig. 442). Then the area under the curve of f between a and b is approximated by n trapezoids of areas

$$\tfrac{1}{2}[f(a) + f(x_1)]h, \qquad \tfrac{1}{2}[f(x_1) + f(x_2)]h, \qquad \cdots, \qquad \tfrac{1}{2}[f(x_{n-1}) + f(b)]h.$$

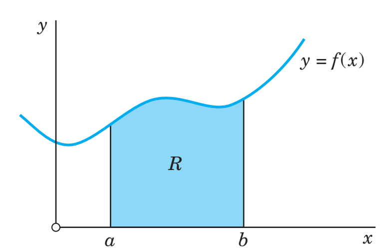


Fig. 440. Geometric interpretation of a definite integral

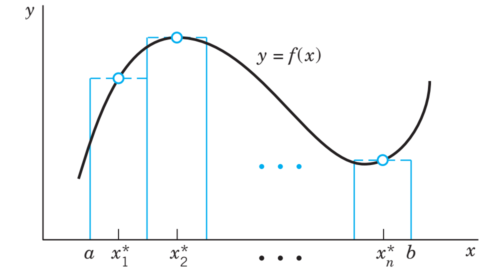


Fig. 441. Rectangular rule

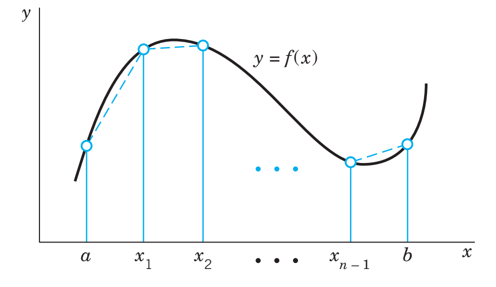


Fig. 442. Trapezoidal rule

By taking their sum we obtain the **trapezoidal rule**

$$J = \int_a^b f(x)\,dx \approx h\left[\frac{1}{2}f(a) + f(x_1) + f(x_2) + \cdots + f(x_{n-1}) + \frac{1}{2}f(b)\right] \qquad \textbf{(2)}$$

where $h = (b - a)/n$, as in (1). The x_j's and a and b are called **nodes**.

EXAMPLE 1 Trapezoidal Rule

Evaluate $J = \int_0^1 e^{-x^2}\,dx$ by means of (2) with $n = 10$.

Note that this integral cannot be evaluated by elementary calculus, but leads to the error function (see Eq. (35), App. 3).

Solution. $J \approx 0.1(0.5 \cdot 1.367879 + 6.778167) = 0.746211$ from Table 19.3. ■

Table 19.3 Computations in Example 1

j	x_j	x_j^2	$e^{-x_j^2}$	
0	0	0	1.000000	
1	0.1	0.01		0.990050
2	0.2	0.04		0.960789
3	0.3	0.09		0.913931
4	0.4	0.16		0.852144
5	0.5	0.25		0.778801
6	0.6	0.36		0.697676
7	0.7	0.49		0.612626
8	0.8	0.64		0.527292
9	0.9	0.81		0.444858
10	1.0	1.00	0.367879	
Sums			1.367879	6.778167

Error Bounds and Estimate for the Trapezoidal Rule

An error estimate for the trapezoidal rule can be derived from (5) in Sec. 19.3 with $n = 1$ by integration as follows. For a single subinterval we have

$$f(x) - p_1(x) = (x - x_0)(x - x_1)\,\frac{f''(t)}{2}$$

with a suitable t depending on x, between x_0 and x_1. Integration over x from $a = x_0$ to $x_1 = x_0 + h$ gives

$$\int_{x_0}^{x_0+h} f(x)\,dx - \frac{h}{2}\,[f(x_0) + f(x_1)] = \int_{x_0}^{x_0+h} (x - x_0)(x - x_0 - h)\,\frac{f''(t(x))}{2}\,dx.$$

Setting $x - x_0 = v$ and applying the mean value theorem of integral calculus, which we can use because $(x - x_0)(x - x_0 - h)$ does not change sign, we find that the right side equals

$$(3^*) \qquad \int_0^h v(v-h)\,dv\,\frac{f''(\tilde{t})}{2} = \left(\frac{h^3}{3} - \frac{h^3}{2}\right)\frac{f''(\tilde{t})}{2} = -\frac{h^3}{12}\,f''(\tilde{t})$$

where $\tilde{t}$ is a (suitable, unknown) value between x_0 and x_1. This is the error for the trapezoidal rule with $n = 1$, often called the **local error**.

Hence the **error** ϵ of (2) with any n is the sum of such contributions from the n subintervals; since $h = (b - a)/n$, $nh^3 = n(b - a)^3/n^3$, and $(b - a)^2 = n^2h^2$, we obtain

$$\textbf{(3)} \qquad \epsilon = -\frac{(b-a)^3}{12n^2}\,f''(\hat{t}) = -\frac{b-a}{12}\,h^2 f''(\hat{t})$$

with (suitable, unknown) $\hat{t}$ between a and b.

Because of (3) the trapezoidal rule (2) is also written

$$(2^*) \qquad J = \int_a^b f(x)\,dx \approx h\left[\frac{1}{2}f(a) + f(x_1) + \cdots + f(x_{n-1}) + \frac{1}{2}f(b)\right] - \frac{b-a}{12}\,h^2 f''(\hat{t}).$$

Error Bounds are now obtained by taking the largest value for f'', say, M_2, and the smallest value, M_2^*, in the interval of integration. Then (3) gives (note that K is negative)

$$\textbf{(4)} \qquad KM_2 \leqq \epsilon \leqq KM_2^* \qquad \text{where} \qquad K = -\frac{(b-a)^3}{12n^2} = -\frac{b-a}{12}\,h^2.$$

Error Estimation by Halving $\boldsymbol{h}$ is advisable if f'' is very complicated or unknown, for instance, in the case of experimental data. Then we may apply the Error Principle of Sec. 19.1. That is, we calculate by (2), first with h, obtaining, say, $J = J_h + \epsilon_h$, and then with $\frac{1}{2}h$, obtaining $J = J_{h/2} + \epsilon_{h/2}$. Now if we replace h^2 in (3) with $(\frac{1}{2}h)^2$, the error is multiplied by $\frac{1}{4}$. Hence $\epsilon_{h/2} \approx \frac{1}{4}\epsilon_h$ (not exactly because $\hat{t}$ may differ). Together, $J_{h/2} + \epsilon_{h/2} = J_h + \epsilon_h \approx J_h + 4\epsilon_{h/2}$. Thus $J_{h/2} - J_h = (4 - 1)\epsilon_{h/2}$. Division by 3 gives the error formula for $J_{h/2}$

$$\textbf{(5)} \qquad \epsilon_{h/2} \approx \tfrac{1}{3}\,(J_{h/2} - J_h).$$

EXAMPLE 2 Error Estimation for the Trapezoidal Rule by (4) and (5)

Estimate the error of the approximate value in Example 1 by (4) and (5).

Solution. **(A)** ***Error bounds by*** **(4).** By differentiation, $f''(x) = 2(2x^2 - 1)e^{-x^2}$. Also, $f'''(x) > 0$ if $0 < x < 1$, so that the minimum and maximum occur at the ends of the interval. We compute $M_2 = f''(1) = 0.735759$ and $M_2^* = f''(0) = -2$. Furthermore, $K = -1/1200$, and (4) gives

$$-0.000614 \leqq \epsilon \leqq 0.001667.$$

Hence the exact value of J must lie between

$$0.746211 - 0.000614 = 0.745597 \qquad \text{and} \qquad 0.746211 + 0.001667 = 0.747878.$$

Actually, $J = 0.746824$, exact to 6D.

(B) ***Error estimate by*** **(5).** $J_h = 0.746211$ in Example 1. Also,

$$J_{h/2} = 0.05\left[\sum_{j=1}^{19} e^{-(j/20)^2} + \frac{1}{2}(1 + 0.367879)\right] = 0.746671.$$

Hence $\epsilon_{h/2} = \frac{1}{3}(J_{h/2} - J_h) = 0.000153$ and $J_{h/2} + \epsilon_{h/2} = 0.746824$, exact to 6D.

Simpson's Rule of Integration

Piecewise constant approximation of f led to the rectangular rule (1), piecewise linear approximation to the trapezoidal rule (2), and piecewise quadratic approximation will lead to Simpson's rule, which is of great practical importance because it is sufficiently accurate for most problems, but still sufficiently simple.

To derive Simpson's rule, we divide the interval of integration $a \leqq x \leqq b$ into an ***even number*** of equal subintervals, say, into $n = 2m$ subintervals of length $h = (b - a)/(2m)$, with endpoints $x_0 (= a), x_1, \cdots, x_{2m-1}, x_{2m} (= b)$; see Fig. 443. We now take the first two subintervals and approximate $f(x)$ in the interval $x_0 \leqq x \leqq x_2 = x_0 + 2h$ by the Lagrange polynomial $p_2(x)$ through $(x_0, f_0), (x_1, f_1), (x_2, f_2)$, where $f_j = f(x_j)$. From (3) in Sec. 19.3 we obtain

$$(6) \quad p_2(x) = \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)} f_0 + \frac{(x - x_0)(x - x_2)}{(x_1 - x_0)(x_1 - x_2)} f_1 + \frac{(x - x_0)(x - x_1)}{(x_2 - x_0)(x_2 - x_1)} f_2.$$

The denominators in (6) are $2h^2$, $-h^2$, and $2h^2$, respectively. Setting $s = (x - x_1)/h$, we have

$$x - x_1 = sh, \qquad x - x_0 = x - (x_1 - h) = (s + 1)h$$
$$x - x_2 = x - (x_1 + h) = (s - 1)h$$

and we obtain

$$p_2(x) = \tfrac{1}{2}s(s - 1)f_0 - (s + 1)(s - 1)f_1 + \tfrac{1}{2}(s + 1)sf_2.$$

We now integrate with respect to x from x_0 to x_2. This corresponds to integrating with respect to s from -1 to 1. Since $dx = h\,ds$, the result is

$$(7^*) \qquad \int_{x_0}^{x_2} f(x)\,dx \approx \int_{x_0}^{x_2} p_2(x)\,dx = h\left(\frac{1}{3} f_0 + \frac{4}{3} f_1 + \frac{1}{3} f_2\right).$$

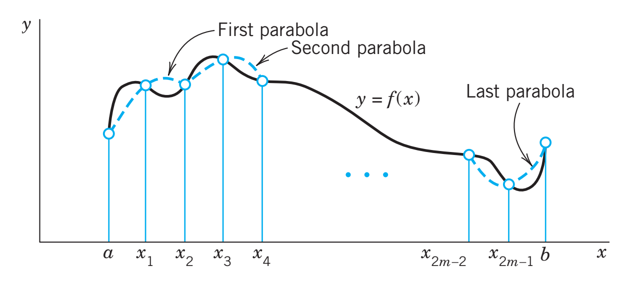


Fig. 443. Simpson's rule

A similar formula holds for the next two subintervals from x_2 to x_4, and so on. By summing all these m formulas we obtain **Simpson's rule**[4]

$$\text{(7)} \qquad \int_a^b f(x)\,dx \approx \frac{h}{3}\,(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m}),$$

where $h = (b - a)/(2m)$ and $f_j = f(x_j)$. Table 19.4 shows an algorithm for Simpson's rule.

Table 19.4 Simpson's Rule of Integration

ALGORITHM SIMPSON $(a, b, m, f_0, f_1, \cdots, f_{2m})$

This algorithm computes the integral $J = \int_a^b f(x)\,dx$ from given values $f_j = f(x_j)$ at equidistant $x_0 = a,\ x_1 = x_0 + h, \cdots, x_{2m} = x_0 + 2mh = b$ by Simpson's rule (7), where $h = (b - a)/(2m)$.

INPUT: $a, b, m, f_0, \cdots, f_{2m}$

OUTPUT: Approximate value $\tilde{J}$ of J

Compute $s_0 = f_0 + f_{2m}$

$s_1 = f_1 + f_3 + \cdots + f_{2m-1}$

$s_2 = f_2 + f_4 + \cdots + f_{2m-2}$

$h = (b - a)/2m$

$\tilde{J} = \dfrac{h}{3}\,(s_0 + 4s_1 + 2s_2)$

OUTPUT $\tilde{J}$. Stop.

End SIMPSON

Error of Simpson's Rule (7). If the fourth derivative $f^{(4)}$ exists and is continuous on $a \leqq x \leqq b$, the **error** of (7), call it ϵ_S, is

$$\text{(8)} \qquad \epsilon_S = -\frac{(b - a)^5}{180\,(2m)^4}\,f^{(4)}(\hat{t}) = -\frac{b - a}{180}\,h^4 f^{(4)}(\hat{t});$$

here $\hat{t}$ is a suitable unknown value between a and b. This is obtained similarly to (3). With this we may also write Simpson's rule (7) as

$$\text{(7**)} \qquad \int_a^b f(x)\,dx = \frac{h}{3}\,(f_0 + 4f_1 + \cdots + f_{2m}) - \frac{b - a}{180}\,h^4 f^{(4)}(\hat{t}).$$

[4]THOMAS SIMPSON (1710–1761), self-taught English mathematician, author of several popular textbooks. Simpson's rule was used much earlier by Torricelli, Gregory (in 1668), and Newton (in 1676).

Error Bounds. By taking for $f^{(4)}$ in (8) the maximum M_4 and minimum M_4^* on the interval of integration we obtain from (8) the error bounds (note that C is negative)

$$\textbf{(9)} \qquad CM_4 \leqq \epsilon_S \leqq CM_4^* \qquad \text{where} \qquad C = -\frac{(b-a)^5}{180(2m)^4} = -\frac{b-a}{180}h^4.$$

Degree of Precision (DP) *of an integration formula.* This is the maximum degree of arbitrary polynomials for which the formula gives exact values of integrals over any intervals.

Hence for the trapezoidal rule,

$$\text{DP} = 1$$

because we approximate the curve of f by portions of straight lines (linear polynomials).

For Simpson's rule we might expect DP = 2 (why?). Actually,

$$\text{DP} = 3$$

by (9) because $f^{(4)}$ is identically zero for a cubic polynomial. This makes Simpson's rule sufficiently accurate for most practical problems and accounts for its popularity.

Numeric Stability *with respect to rounding* is another important property of Simpson's rule. Indeed, for the sum of the roundoff errors ϵ_j of the $2m + 1$ values f_j in (7) we obtain, since $h = (b - a)/2m$,

$$\frac{h}{3}\left|\epsilon_0 + 4\epsilon_1 + \cdots + \epsilon_{2m}\right| \leqq \frac{b-a}{3.2m}\,6mu = (b-a)u$$

where u is the rounding unit ($u = \frac{1}{2} \cdot 10^{-6}$ if we round off to 6D; see Sec. 19.1). Also $6 = 1 + 4 + 1$ is the sum of the coefficients for a pair of intervals in (7); take $m = 1$ in (7) to see this. The bound $(b - a)u$ is independent of m, so that it cannot increase with increasing m, that is, with decreasing h. This proves stability. ■

Newton–Cotes Formulas. We mention that the trapezoidal and Simpson rules are special *closed Newton–Cotes formulas*, that is, integration formulas in which $f(x)$ is interpolated at equally spaced nodes by a polynomial of degree n ($n = 1$ for trapezoidal, $n = 2$ for Simpson), and **closed** means that a and b are nodes ($a = x_0$, $b = x_n$). $n = 3$ and higher n are used occasionally. From $n = 8$ on, some of the coefficients become negative, so that a positive f_j could make a negative contribution to an integral, which is absurd. For more on this topic see Ref. [E25] in App. 1.

EXAMPLE 3 Simpson's Rule. Error Estimate

Evaluate $J = \int_0^1 e^{-x^2} dx$ by Simpson's rule with $2m = 10$ and estimate the error.

Solution. Since $h = 0.1$, Table 19.5 gives

$$J \approx \frac{0.1}{3}(1.367879 + 4 \cdot 3.740266 + 2 \cdot 3.037901) = 0.746825.$$

Estimate of error. Differentiation gives $f^{(4)}(x) = 4(4x^4 - 12x^2 + 3)e^{-x^2}$. By considering the derivative $f^{(5)}$ of $f^{(4)}$ we find that the largest value of $f^{(4)}$ in the interval of integration occurs at 0 and the smallest value at $x^* = (2.5 - 0.5\sqrt{10})^{1/2}$. Computation gives the values $M_4 = f^{(4)}(0) = 12$ and $M_4^* = f^{(4)}(x^*) = -7.419$. Since $2m = 10$ and $b - a = 1$, we obtain $C = -1/1800000 = -0.00000056$. Therefore, from (9),

$$-0.000007 \leqq \epsilon_S \leqq 0.000005.$$

Hence J must lie between $0.746825 - 0.000007 = 0.746818$ and $0.746825 + 0.000005 = 0.746830$, so that at least four digits of our approximate value are exact. Actually, the value 0.746825 is exact to 5D because $J = 0.746824$ (exact to 6D).

Thus our result is much better than that in Example 1 obtained by the trapezoidal rule, whereas the number of operations is nearly the same in both cases. ■

Table 19.5 Computations in Example 3

j	x_j	x_j^2	$e^{-x_j^2}$		
0	0	0	1.000000		
1	0.1	0.01		0.990050	
2	0.2	0.04			0.960789
3	0.3	0.09		0.913931	
4	0.4	0.16			0.852144
5	0.5	0.25		0.778801	
6	0.6	0.36			0.697676
7	0.7	0.49		0.612626	
8	0.8	0.64			0.527292
9	0.9	0.81		0.444858	
10	1.0	1.00	0.367879		
Sums			1.367879	3.740266	3.037901

Instead of picking an $n = 2m$ and then estimating the error by (9), as in Example 3, it is better to require an accuracy (e.g., 6D) and then determine $n = 2m$ from (9).

EXAMPLE 4 Determination of $n = 2m$ in Simpson's Rule from the Required Accuracy

What n should we choose in Example 3 to get 6D-accuracy?

Solution. Using $M_4 = 12$ (which is bigger in absolute value than M_4^*, we get from (9), with $b - a = 1$ and the required accuracy,

$$|CM_4| = \frac{12}{180(2m)^4} = \frac{1}{2} \cdot 10^{-6}, \qquad \text{thus} \qquad m = \left[\frac{2 \cdot 10^6 \cdot 12}{180 \cdot 2^4}\right]^{1/4} = 9.55.$$

Hence we should choose $n = 2m = 20$. Do the computation, which parallels that in Example 3.

Note that the error bounds in (4) or (9) may sometimes be loose, so that in such a case a smaller $n = 2m$ may already suffice. ■

Error Estimation for Simpson's Rule by Halving h. The idea is the same as in (5) and gives

$$\epsilon_{h/2} \approx \tfrac{1}{15}(J_{h/2} - J_h). \tag{10}$$

J_h is obtained by using h and $J_{h/2}$ by using $\frac{1}{2}h$, and $\epsilon_{h/2}$ is the error of $J_{h/2}$.

Derivation. In (5) we had $\frac{1}{3}$ as the reciprocal of $3 = 4 - 1$ and $\frac{1}{4} = (\frac{1}{2})^2$ resulted from h^2 in (3) by replacing h with $\frac{1}{2}h$. In (10) we have $\frac{1}{15}$ as the reciprocal of $15 = 16 - 1$ and $\frac{1}{16} = (\frac{1}{2})^4$ results from h^4 in (8) by replacing h with $\frac{1}{2}h$.

EXAMPLE 5 Error Estimation for Simpson's Rule by Halving

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from 0 to 2 with $h = 1$ and apply (10).

Solution. The exact 5D-value of the integral is $J = 1.25953$. Simpson's rule gives

$$J_h = \tfrac{1}{3}[f(0) + 4f(1) + f(2)] = \tfrac{1}{3}(0 + 4 \cdot 0.555360 + 0) = 0.740480,$$

$$J_{h/2} = \tfrac{1}{6}\,[f(0) + 4f(\tfrac{1}{2}) + 2f(1) + 4f(\tfrac{3}{2}) + f(2)]$$

$$= \tfrac{1}{6}[0 + 4 \cdot 0.045351 + 2 \cdot 0.555361 + 4 \cdot 1.521579 + 0] = 1.22974.$$

Hence (10) gives $\epsilon_{h/2} = \frac{1}{15}(1.22974 - 0.74048) = 0.032617$ and thus $J \approx J_{h/2} + \epsilon_{h/2} = 1.26236$, with an error -0.00283 which is less in absolute value than $\frac{1}{10}$ of the error 0.02979 of $J_{h/2}$. Hence the use of (10) was well worthwhile. ■

Adaptive Integration

The idea is to adapt step h to the variability of $f(x)$. That is, where f varies but little, we can proceed in large steps without causing a substantial error in the integral, but where f varies rapidly, we have to take small steps in order to stay everywhere close enough to the curve of f.

Changing h is done systematically, usually by halving h, and automatically (not "by hand") depending on the size of the (estimated) error over a subinterval. The subinterval is halved if the corresponding error is still too large, that is, larger than a given **tolerance** TOL (maximum admissible absolute error), or is not halved if the error is less than or equal to TOL (or doubled if the error is very small).

Adapting is one of the techniques typical of modern software. In connection with integration it can be applied to various methods. We explain it here for Simpson's rule. In Table 19.6 an asterisk means that for that subinterval, TOL has been reached.

EXAMPLE 6 Adaptive Integration with Simpson's Rule

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 by adaptive integration and with Simpson's rule and TOL[0, 2] = 0.0002.

Solution. Table 19.6 shows the calculations. Figure 444 shows the integrand $f(x)$ and the adapted intervals used. The first two intervals ([0, 0.5], [0.5, 1.0]) have length 0.5, hence $h = 0.25$ [because we use $2m = 2$ subintervals in Simpson's rule (7**)]. The next two intervals ([1.00, 1.25], [1.25, 1.50]) have length 0.25 (hence $h = 0.125$) and the last four intervals have length 0.125. *Sample computations.* For 0.740480 see Example 5. Formula (10) gives $(0.123716 - 0.122794)/15 = 0.000061$. Note that 0.123716 refers to [0, 0.5] and [0.5, 1], so that we must subtract the value corresponding to [0, 1] in the line before. Etc. TOL[0, 2] = 0.0002 gives 0.0001 for subintervals of length 1, 0.00005 for length 0.5, etc. The value of the integral obtained is the sum of the values marked by an asterisk (for which the error estimate has become less than TOL). This gives

$$J \approx 0.123716 + 0.528895 + 0.388263 + 0.218483 = 1.25936.$$

The exact 5D-value is $J = 1.25953$. Hence the error is 0.00017. This is about 1/200 of the absolute value of that in Example 5. Our more extensive computation has produced a much better result. ■

Table 19.6 Computations in Example 6

Interval	Integral	Error (10)	TOL	Comment
[0, 2]	0.740480		0.0002	
[0, 1]	0.122794			
[1, 2]	1.10695			
	Sum = 1.22974	0.032617	0.0002	Divide further
[0.0, 0.5]	0.004782			
[0.5, 1.0]	0.118934			
	Sum = 0.123716*	0.000061	0.0001	TOL reached
[1.0, 1.5]	0.528176			
[1.5, 2.0]	0.605821			
	Sum = 1.13300	0.001803	0.0001	Divide further
[1.00, 1.25]	0.200544			
[1.25, 1.50]	0.328351			
	Sum = 0.528895*	0.000048	0.00005	TOL reached
[1.50, 1.75]	0.388235			
[1.75, 2.00]	0.218457			
	Sum = 0.606692	0.000058	0.00005	Divide further
[1.500, 1.625]	0.196244			
[1.625, 1.750]	0.192019			
	Sum = 0.388263*	0.000002	0.000025	TOL reached
[1.750, 1.875]	0.153405			
[1.875, 2.000]	0.065078			
	Sum = 0.218483*	0.000002	0.000025	TOL reached

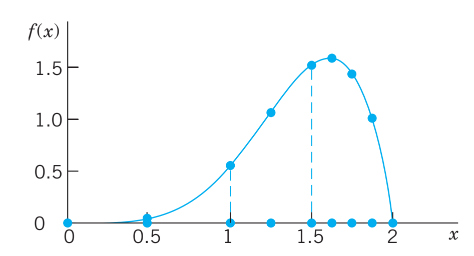


Fig. 444. Adaptive integration in Example 6

Gauss Integration Formulas
Maximum Degree of Precision

Our integration formulas discussed so far use function values at *predetermined* (equidistant) x-values (nodes) and give exact results for polynomials not exceeding a

certain degree [called the *degree of precision*; see after (9)]. But we can get much more accurate integration formulas as follows. We set

$$\int_{-1}^{1} f(t)\,dt \approx \sum_{j=1}^{n} A_j f_j \qquad [f_j = f(t_j)] \tag{11}$$

with fixed n, and $t = \pm 1$ obtained from $x = a, b$ by setting $x = \frac{1}{2}[a(t-1) + b(t+1)]$. Then we determine the n coefficients $A_1, \cdots, A_n$ and n nodes $t_1, \cdots, t_n$ so that (11) gives exact results for polynomials of degree k as high as possible. Since $n + n = 2n$ is the number of coefficients of a polynomial of degree $2n - 1$, it follows that $k \leqq 2n - 1$.

Gauss has shown that exactness for polynomials of degree not exceeding $2n - 1$ (instead of $n - 1$ for predetermined nodes) can be attained, and he has given the location of the t_j (= the jth zero of the Legendre polynomial P_n in Sec. 5.3) and the coefficients A_j which depend on n but not on $f(t)$, and are obtained by using Lagrange's interpolation polynomial, as shown in Ref. [E5] listed in App. 1. With these t_j and A_j, formula (11) is called a **Gauss integration formula** or *Gauss quadrature formula*. Its degree of precision is $2n - 1$, as just explained. Table 19.7 gives the values needed for $n = 2, \cdots, 5$. (For larger n, see pp. 916–919 of Ref. [GenRef1] in App. 1.)

Table 19.7 Gauss Integration: Nodes t_j and Coefficients A_j

n	Nodes t_j	Coefficients A_j	Degree of Precision
2	−0.5773502692	1	3
	0.5773502692	1	
3	−0.7745966692	0.5555555556	5
	0	0.8888888889	
	0.7745966692	0.5555555556	
4	−0.8611363116	0.3478548451	7
	−0.3399810436	0.6521451549	
	0.3399810436	0.6521451549	
	0.8611363116	0.3478548451	
5	−0.9061798459	0.2369268851	9
	−0.5384693101	0.4786286705	
	0	0.5688888889	
	0.5384693101	0.4786286705	
	0.9061798459	0.2369268851	

EXAMPLE 7 Gauss Integration Formula with $n = 3$

Evaluate the integral in Example 3 by the Gauss integration formula (11) with $n = 3$.

Solution. We have to convert our integral from 0 to 1 into an integral from -1 to 1. We set $x = \frac{1}{2}(t + 1)$. Then $dx = \frac{1}{2}\,dt$, and (11) with $n = 3$ and the above values of the nodes and the coefficients yields

$$\int_0^1 \exp(-x^2)\,dx = \frac{1}{2}\int_{-1}^{1} \exp\left(-\frac{1}{4}(t+1)^2\right)dt$$

$$\approx \frac{1}{2}\left[\frac{5}{9}\exp\left(-\frac{1}{4}\left(1-\sqrt{\frac{3}{5}}\right)^2\right) + \frac{8}{9}\exp\left(-\frac{1}{4}\right) + \frac{5}{9}\exp\left(-\frac{1}{4}\left(1+\sqrt{\frac{3}{5}}\right)^2\right)\right] = 0.746815$$

(exact to 6D: 0.746825), which is almost as accurate as the Simpson result obtained in Example 3 with a much larger number of arithmetic operations. With 3 function values (as in this example) and Simpson's rule we would get $\frac{1}{6}(1 + 4e^{-0.25} + e^{-1}) = 0.747180$, with an error over 30 times that of the Gauss integration. ■

EXAMPLE 8 Gauss Integration Formula with $n = 4$ and 5

Integrate $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 by Gauss. Compare with the adaptive integration in Example 6 and comment.

Solution. $x = t + 1$ gives $f(t) = \frac{1}{4}\pi(t+1)^4 \cos(\frac{1}{4}\pi(t+1))$, as needed in (11). For $n = 4$ we calculate (6S)

$$J \approx A_1 f_1 + \cdots + A_4 f_4 = A_1(f_1 + f_4) + A_2(f_2 + f_3)$$

$$= 0.347855(0.000290309 + 1.02570) + 0.652145(0.129464 + 1.25459) = 1.25950.$$

The error is 0.00003 because $J = 1.25953$ (6S). Calculating with 10S and $n = 4$ gives the same result; so the error is due to the formula, not rounding. For $n = 5$ and 10S we get $J \approx 1.259526185$, too large by the amount 0.000000250 because $J = 1.259525935$ (10S). The accuracy is impressive, particularly if we compare the amount of work with that in Example 6. ■

Gauss integration is of considerable practical importance. Whenever the integrand f is given by a formula (not just by a table of numbers) or when experimental measurements can be set at times t_j (or whatever t represents) shown in Table 19.7 or in Ref. [GenRef1], then the great accuracy of Gauss integration outweighs the disadvantage of the complicated t_j and A_j (which may have to be stored). Also, Gauss coefficients A_j are positive for all n, in contrast with some of the Newton–Cotes coefficients for larger n.

Of course, there are frequent applications with equally spaced nodes, so that Gauss integration does not apply (or has no great advantage if one first has to get the t_j in (11) by interpolation).

Since the endpoints -1 and 1 of the interval of integration in (11) are not zeros of P_n, they do not occur among $t_0, \cdots, t_n$, and the Gauss formula (11) is called, therefore, an **open formula**, in contrast with a **closed formula**, in which the endpoints of the interval of integration are t_0 and t_n. [For example, (2) and (7) are closed formulas.]

Numeric Differentiation

Numeric differentiation is the computation of values of the derivative of a function f from given values of f. Numeric differentiation should be avoided whenever possible. Whereas *integration* is a smoothing process and is not very sensitive to small inaccuracies in function values, *differentiation* tends to make matters rough and generally gives values of f' that are much less accurate than those of f. The difficulty with differentiation is tied in with the definition of the derivative, which is the limit of the difference quotient, and, in that quotient, you usually have the difference of a large quantity divided by a small quantity. This can cause numerical instability. While being aware of this caveat, we must still develop basic differentiation formulas for use in numeric solutions of differential equations.

We use the notations $f_j' = f'(x_j)$, $f_j'' = f''(x_j)$, etc., and may obtain rough approximation formulas for derivatives by remembering that

$$f'(x) = \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}.$$

This suggests

$$f'_{1/2} \approx \frac{\delta f_{1/2}}{h} = \frac{f_1 - f_0}{h}. \tag{12}$$

Similarly, for the second derivative we obtain

$$f''_1 \approx \frac{\delta^2 f_1}{h^2} = \frac{f_2 - 2f_1 + f_0}{h^2}, \qquad \text{etc.} \tag{13}$$

More accurate approximations are obtained by differentiating suitable Lagrange polynomials. Differentiating (6) and remembering that the denominators in (6) are $2h^2$, $-h^2$, $2h^2$, we have

$$f'(x) \approx p'_2(x) = \frac{2x - x_1 - x_2}{2h^2} f_0 - \frac{2x - x_0 - x_2}{h^2} f_1 + \frac{2x - x_0 - x_1}{2h^2} f_2.$$

Evaluating this at x_0, x_1, x_2, we obtain the "three-point formulas"

$$\text{(14)} \qquad \begin{aligned} &\text{(a)} \quad f'_0 \approx \frac{1}{2h}(-3f_0 + 4f_1 - f_2), \\ &\text{(b)} \quad f'_1 \approx \frac{1}{2h}(-f_0 + f_2), \\ &\text{(c)} \quad f'_2 \approx \frac{1}{2h}(f_0 - 4f_1 + 3f_2). \end{aligned}$$

Applying the same idea to the Lagrange polynomial $p_4(x)$, we obtain similar formulas, in particular,

$$f'_2 \approx \frac{1}{12h}(f_0 - 8f_1 + 8f_3 - f_4). \tag{15}$$

Some examples and further formulas are included in the problem set as well as in Ref. [E5] listed in App. 1.

PROBLEM SET 19.5

1–6 RECTANGULAR AND TRAPEZOIDAL RULES

1. **Rectangular rule.** Evaluate the integral in Example 1 by the rectangular rule (1) with subintervals of length 0.1. Compare with Example 1. (6S-exact: 0.746824)
2. **Bounds for (1).** Derive a formula for lower and upper bounds for the rectangular rule. Apply it to Prob. 1.
3. **Trapezoidal rule.** To get a feel for increase in accuracy, integrate x^2 from 0 to 1 by (2) with $h = 1, 0.5, 0.25, 0.1$.
4. **Error estimation by halfing.** Integrate $f(x) = x^4$ from 0 to 1 by (2) with $h = 1$, $h = 0.5$, $h = 0.25$ and estimate the error for $h = 0.5$ and $h = 0.25$ by (5).
5. **Error estimation.** Do the tasks in Prob. 4 for $f(x) = \sin \frac{1}{2}\pi x$.

6. Stability. Prove that the trapezoidal rule is stable with respect to rounding.

7–15 SIMPSON'S RULE

Evaluate the integrals $A = \int_1^2 \frac{dx}{x}$, $B = \int_0^{0.4} xe^{-x^2}\, dx$, $J = \int_0^1 \frac{dx}{1 + x^2}$ by Simpson's rule with $2m$ as indicated, and compare with the exact value known from calculus.

7. A, $2m = 4$

8. A, $2m = 10$

9. B, $2m = 4$

10. B, $2m = 10$

11. J, $2m = 4$

12. J, $2m = 10$

13. Error estimate. Compute the integral J by Simpson's rule with $2m = 8$ and use the value and that in Prob. 11 to estimate the error by (10).

14. Error bounds and estimate. Integrate e^{-x} from 0 to 2 by (7) with $h = 1$ and with $h = 0.5$. Give error bounds for the $h = 0.5$ value and an error estimate by (10).

15. Given TOL. Find the smallest n in computing A (see Probs. 7 and 8) such that 5S-accuracy is guaranteed **(a)** by (4) in the use of (2), **(b)** by (9) in the use of (7).

16–21 NONELEMENTARY INTEGRALS

The following integrals cannot be evaluated by the usual methods of calculus. Evaluate them as indicated. Compare your value with that possibly given by your CAS. Si(x) is the sine integral. S(x) and C(x) are the Fresnel integrals. See App. A3.1. They occur in optics.

$$\text{Si}(x) = \int_0^x \frac{\sin x^*}{x^*}\, dx^*,$$

$$\text{S}(x) = \int_0^x \sin(x^{*2})\, dx^*, \quad \text{C}(x) = \int_0^x \cos(x^{*2})\, dx^*$$

16. Si(1) by (2), $n = 5$, $n = 10$, and apply (5).

17. Si(1) by (7), $2m = 2$, $2m = 4$

18. Obtain a better value in Prob. 17. *Hint.* Use (10).

19. Si(1) by (7), $2m = 10$

20. S(1.25) by (7), $2m = 10$

21. C(1.25) by (7), $2m = 10$

22–25 GAUSS INTEGRATION

Integrate by (11) with $n = 5$:

22. $\cos x$ from 0 to $\frac{1}{2}\pi$

23. xe^{-x} from 0 to 1

24. $\sin(x^2)$ from 0 to 1.25

25. $\exp(-x^2)$ from 0 to 1

26. TEAM PROJECT. Romberg Integration (W. Romberg, *Norske Videnskab. Trondheim, Fφrh.* 28, Nr. 7, 1955). This method uses the trapezoidal rule and gains precision stepwise by halving h and adding an error estimate. Do this for the integral of $f(x) = e^{-x}$ from $x = 0$ to $x = 2$ with TOL $= 10^{-3}$, as follows.

Step 1. Apply the trapezoidal rule (2) with $h = 2$ (hence $n = 1$) to get an approximation J_{11}. Halve h and use (2) to get J_{21} and an error estimate

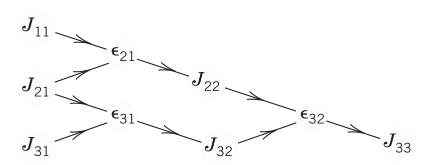
$$\epsilon_{21} = \frac{1}{2^2 - 1}(J_{21} - J_{11}).$$

If $|\epsilon_{21}| \leqq$ TOL, stop. The result is $J_{22} = J_{21} + \epsilon_{21}$.

Step 2. Show that $\epsilon_{21} = -0.066596$, hence $|\epsilon_{21}| >$ TOL and go on. Use (2) with $h/4$ to get J_{31} and add to it the error estimate $\epsilon_{31} = \frac{1}{3}(J_{31} - J_{21})$ to get the better $J_{32} = J_{31} + \epsilon_{31}$. Calculate

$$\epsilon_{32} = \frac{1}{2^4 - 1}(J_{32} - J_{22}) = \frac{1}{15}(J_{32} - J_{22}).$$

If $|\epsilon_{32}| \leqq$ TOL, stop. The result is $J_{33} = J_{32} + \epsilon_{32}$. (Why does $2^4 = 16$ come in?) Show that we obtain $\epsilon_{32} = -0.000266$, so that we can stop. Arrange your J- and ϵ-values in a kind of "difference table."



If $|\epsilon_{32}|$ were greater than TOL, you would have to go on and calculate in the next step J_{41} from (2) with $h = \frac{1}{4}$; then

$$J_{42} = J_{41} + \epsilon_{41} \quad \text{with} \quad \epsilon_{41} = \tfrac{1}{3}(J_{41} - J_{31})$$
$$J_{43} = J_{42} + \epsilon_{42} \quad \text{with} \quad \epsilon_{42} = \tfrac{1}{15}(J_{42} - J_{32})$$
$$J_{44} = J_{43} + \epsilon_{43} \quad \text{with} \quad \epsilon_{43} = \tfrac{1}{63}(J_{43} - J_{33})$$

where $63 = 2^6 - 1$. (How does this come in?)

Apply the Romberg method to the integral of $f(x) = \frac{1}{4}\pi x^4 \cos \frac{1}{4}\pi x$ from $x = 0$ to 2 with TOL $= 10^{-4}$.

27–30 DIFFERENTIATION

27. Consider $f(x) = x^4$ for $x_0 = 0, x_1 = 0.2, x_2 = 0.4, x_3 = 0.6, x_4 = 0.8$. Calculate f_2' from (14a), (14b), (14c), (15). Determine the errors. Compare and comment.

28. A "**four-point formula**" for the derivative is

$$f_2' \approx \frac{1}{6h}\,(-2f_1 - 3f_2 + 6f_3 - f_4).$$

Apply it to $f(x) = x^4$ with $x_1, \cdots, x_4$ as in Prob. 27, determine the error, and compare it with that in the case of (15).

29. The derivative $f'(x)$ can also be approximated in terms of first-order and higher order differences (see Sec. 19.3):

$$f'(x_0) \approx \frac{1}{h}\left(\Delta f_0 - \frac{1}{2}\Delta^2 f_0 + \frac{1}{3}\Delta^3 f_0 - \frac{1}{4}\Delta^4 f_0 + - \cdots\right).$$

Compute $f'(0.4)$ in Prob. 27 from this formula, using differences up to and including first order, second order, third order, fourth order.

30. Derive the formula in Prob. 29 from (14) in Sec. 19.3.

CHAPTER 19 REVIEW QUESTIONS AND PROBLEMS

1. What is a numeric method? How has the computer influenced numerics?

2. What is an error? A relative error? An error bound?

3. Why are roundoff errors important? State the rounding rules.

4. What is an algorithm? Which of its properties are important in software implementation?

5. What do you know about stability?

6. Why is the selection of a *good* method at least as important on a large computer as it is on a small one?

7. Can the Newton (–Raphson) method diverge? Is it fast? Same questions for the bisection method.

8. What is fixed-point iteration?

9. What is the advantage of Newton's interpolation formulas over Lagrange's?

10. What is spline interpolation? Its advantage over polynomial interpolation?

11. List and compare the integration methods we have discussed.

12. How did we use an interpolation polynomial in deriving Simpson's rule?

13. What is adaptive integration? Why is it useful?

14. In what sense is Gauss integration optimal?

15. How did we obtain formulas for numeric differentiation?

16. Write -46.9028104, 0.000317399, $54/7$, $-890/3$ in floating-point form with 5S (5 significant digits, properly rounded).

17. Compute $(5.346 - 3.644)/(3.444 - 3.055)$ as given and then rounded stepwise to 3S, 2S, 1S. Comment. ("Stepwise" means rounding the rounded numbers, not the given ones.)

18. Compute $0.38755/(5.6815 - 0.38419)$ as given and then rounded stepwise to 4S, 3S, 2S, 1S. Comment.

19. Let 19.1 and 25.84 be correctly rounded. Find the shortest interval in which the sum s of the true (unrounded) numbers must lie.

20. Do the same task as in Prob. 19 for the difference $3.2 - 6.29$.

21. What is the relative error of $n\tilde{a}$ in terms of that of $\tilde{a}$?

22. Show that the relative error of $\tilde{a}^2$ is about twice that of $\tilde{a}$.

23. Solve $x^2 - 40x + 2 = 0$ in two ways (cf. Sec. 19.1). Use 4S-arithmetic.

24. Solve $x^2 - 100x + 1 = 0$. Use 5S-arithmetic.

25. Compute the solution of $x^4 = x + 0.1$ near $x = 0$ by transforming the equation algebraically to the form $x = g(x)$ and starting from $x_0 = 0$.

26. Solve $\cos x = x^2$ by Newton's method, starting from $x = 0.5$.

27. Solve Prob. 25 by bisection (3S-accuracy).

28. Compute $\sinh 0.4$ from $\sinh 0$, $\sinh 0.5 = 0.521$, $\sinh 1.0 = 1.175$ by quadratic interpolation.

29. Find the cubic spline for the data $f(0) = 0$, $f(1) = 0$, $f(2) = 4$, $k_0 = -1$, $k_2 = 5$.

30. Find the cubic spline q and the interpolation polynomial p for the data (0, 0), (1, 1), (2, 6), (3, 10), with $q'(0) = 0$, $q'(3) = 0$ and graph p and q on common axes.

31. Compute the integral of x^3 from 0 to 1 by the trapezoidal rule with $n = 5$. What error bounds are obtained from (4) in Sec. 19.5? What is the actual error of the result?

32. Compute the integral of $\cos(x^2)$ from 0 to 1 by Simpson's rule with $2m = 4$.

33. Solve Prob. 32 by Gauss integration with $n = 3$ and $n = 5$.

34. Compute $f'(0.2)$ for $f(x) = x^3$ using (14b) in Sec. 19.5 with **(a)** $h = 0.2$, **(b)** $h = 0.1$. Compare the accuracy.

35. Compute $f''(0.2)$ for $f(x) = x^3$ using (13) in Sec. 19.5 with **(a)** $h = 0.2$, **(b)** $h = 0.1$.

SUMMARY OF CHAPTER 19
Numerics in General

In this chapter we discussed concepts that are relevant throughout numeric work as a whole and methods of a general nature, as opposed to methods for linear algebra (Chap. 20) or differential equations (Chap. 21).

In scientific computations we use the ***floating-point*** representation of numbers (Sec. 19.1); fixed-point representation is less suitable in most cases.

Numeric methods give approximate values $\tilde{a}$ of quantities. The **error** ϵ of $\tilde{a}$ is

$$(1) \qquad \epsilon = a - \tilde{a} \qquad \text{(Sec. 19.1)}$$

where a is the exact value. The *relative error* of $\tilde{a}$ is ϵ/a. Errors arise from rounding, inaccuracy of measured values, truncation (that is, replacement of integrals by sums, series by partial sums), and so on.

An algorithm is called **numerically stable** if small changes in the initial data give only correspondingly small changes in the final results. Unstable algorithms are generally useless because errors may become so large that results will be very inaccurate. The numeric instability of algorithms must not be confused with the mathematical instability of problems ("*ill-conditioned problems*," Sec. 19.2).

Fixed-point iteration is a method for solving equations $f(x) = 0$ in which the equation is first transformed algebraically to $x = g(x)$, an initial guess x_0 for the solution is made, and then approximations $x_1, x_2, \cdots$, are successively computed by iteration from (see Sec. 19.2)

$$(2) \qquad x_{n+1} = g(x_n) \qquad (n = 0, 1, \cdots).$$

Newton's method for solving equations $f(x) = 0$ is an iteration

$$(3) \qquad x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)} \qquad \text{(Sec. 19.2).}$$

Here x_{n+1} is the x-intercept of the tangent of the curve $y = f(x)$ at the point x_n. This method is of second order (Theorem 2, Sec. 19.2). If we replace f' in (3) by a difference quotient (geometrically: we replace the tangent by a secant), we obtain the **secant method;** see (10) in Sec. 19.2. For the *bisection method* (which converges slowly) and the *method of false position,* see Problem Set 19.2.

Polynomial interpolation means the determination of a polynomial $p_n(x)$ such that $p_n(x_j) = f_j$, where $j = 0, \cdots, n$ and $(x_0, f_0), \cdots, (x_n, f_n)$ are measured or observed values, values of a function, etc. $p_n(x)$ is called an *interpolation polynomial.* For given data, $p_n(x)$ of degree n (or less) is unique. However, it can be written in different forms, notably in **Lagrange's form** (4), Sec. 19.3, or in **Newton's divided difference form** (10), Sec. 19.3, which requires fewer operations. For regularly spaced $x_0, x_1 = x_0 + h, \cdots, x_n = x_0 + nh$ the latter becomes **Newton's forward difference formula** (formula (14) in Sec. 19.3):

$$(4) \qquad f(x) \approx p_n(x) = f_0 + r\,\Delta f_0 + \cdots + \frac{r(r-1)\cdots(r-n+1)}{n!}\,\Delta^n f_0$$

where $r = (x - x_0)/h$ and the forward differences are $\Delta f_j = f_{j+1} - f_j$ and

$$\Delta^k f_j = \Delta^{k-1} f_{j+1} - \Delta^{k-1} f_j \qquad (k = 2, 3, \cdots).$$

A similar formula is *Newton's backward difference interpolation formula* (formula (18) in Sec. 19.3).

Interpolation polynomials may become numerically unstable as n increases, and instead of interpolating and approximating by a single high-degree polynomial it is preferable to use a cubic **spline** $g(x)$, that is, a twice continuously differentiable interpolation function [thus, $g(x_j) = f_j$], which in each subinterval $x_j \leqq x \leqq x_{j+1}$ consists of a cubic polynomial $q_j(x)$; see Sec. 19.4.

Simpson's rule of numeric integration is [see (7), Sec. 19.5]

$$(5) \qquad \int_a^b f(x)\,dx \approx \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \cdots + 2f_{2m-2} + 4f_{2m-1} + f_{2m})$$

with equally spaced nodes $x_j = x_0 + jh, j = 1, \cdots, 2m, h = (b - a)/(2m)$, and $f_j = f(x_j)$. It is simple but accurate enough for many applications. Its degree of precision is DP $= 3$ because the error (8), Sec. 19.5, involves h^4. A more practical error estimate is (10), Sec. 19.5,

$$\epsilon_{h/2} = \tfrac{1}{15}(J_{h/2} - J_h),$$

obtained by first computing with step h, then with step $h/2$, and then taking $\frac{1}{15}$ of the difference of the results.

Simpson's rule is the most important of the **Newton–Cotes formulas**, which are obtained by integrating Lagrange interpolation polynomials, linear ones for the **trapezoidal rule** (2), Sec. 19.5, quadratic for Simpson's rule, cubic for the *three-eights rule* (see the Chap. 19 Review Problems), etc.

Adaptive integration (Sec. 19.5, Example 6) is integration that adjusts (*"adapts"*) the step (automatically) to the variability of $f(x)$.

Romberg integration (Team Project 26, Problem Set 19.5) starts from the trapezoidal rule (2), Sec. 19.5, with $h, h/2, h/4$, etc. and improves results by systematically adding error estimates.

Gauss integration (11), Sec. 19.5, is important because of its great accuracy (DP $= 2n - 1$, compared to Newton–Cotes's DP $= n - 1$ or n). This is achieved by an optimal choice of the nodes, which are not equally spaced; see Table 19.7, Sec. 19.5.

Numeric differentiation is discussed at the end of Sec. 19.5. (Its main application (to differential equations) follows in Chap. 21.)

CHAPTER 20

Numeric Linear Algebra

This chapter deals with two main topics. The first topic is how to solve linear systems of equations numerically. We start with Gauss elimination, which may be familiar to some readers, but this time in an algorithmic setting with partial pivoting. Variants of this method (Doolittle, Crout, Cholesky, Gauss–Jordan) are discussed in Sec. 20.2. All these methods are direct methods, that is, methods of numerics where we know in advance how many steps they will take until they arrive at a solution. However, small pivots and roundoff error magnification may produce nonsensical results, such as in the Gauss method. A shift occurs in Sec. 20.3, where we discuss numeric iteration methods or indirect methods to address our first topic. Here we cannot be totally sure how many steps will be needed to arrive at a good answer. Several factors—such as how far is the starting value from our initial solution, how is the problem structure influencing speed of convergence, how accurate would we like our result to be—determine the outcome of these methods. Moreover, our computation cycle may not converge. Gauss–Seidel iteration and Jacobi iteration are discussed in Sec. 20.3. Section 20.4 is at the heart of addressing the pitfalls of numeric linear algebra. It is concerned with problems that are ill-conditioned. We learn to estimate how "bad" such a problem is by calculating the condition number of its matrix.

The second topic (Secs. 20.6–20.9) is how to solve eigenvalue problems numerically. Eigenvalue problems appear throughout engineering, physics, mathematics, economics, and many areas. For large or very large matrices, determining the eigenvalues is difficult as it involves finding the roots of the characteristic equations, which are high-degree polynomials. As such, there are different approaches to tackling this problem. Some methods, such as Gerschgorin's method and Collatz's method only provide a range in which eigenvalues lie and thus are known as inclusion methods. Others such as tridiagonalization and QR-factorization actually find all the eigenvalues. The area is quite ingeneous and should be fascinating to the reader.

COMMENT. ***This chapter is independent of Chap. 19 and can be studied immediately after Chap. 7 or 8.***

Prerequisite: Secs. 7.1, 7.2, 8.1.
Sections that may be omitted in a shorter course: 20.4, 20.5, 20.9.
References and Answers to Problems: App. 1 Part E, App. 2.

20.1 Linear Systems: Gauss Elimination

The basic method for solving systems of linear equations by Gauss elimination and back substitution was explained in Sec. 7.3. If you covered Sec. 7.3, you may wonder why we cover Gauss elimination again. The reason is that *here we cover Gauss elimination in the*

setting of numerics and introduce new material such as pivoting, row scaling, and operation count. Furthermore, we give an algorithmic representation of Gauss elimination in Table 20.1 that can be readily converted into software. We also show when Gauss elimination runs into difficulties with small pivots and what to do about it. The reader should pay close attention to the material as variants of Gauss elimination are covered in Sec. 20.2 and, furthermore, the general problem of solving linear systems is the focus of the first half of this chapter.

A **linear system of *n* equations** in n *unknowns* $x_1, \cdots, x_n$ is a set of equations $\mathrm{E}_1, \cdots, \mathrm{E}_n$ of the form

$$
(1) \qquad \begin{array}{ll}
\mathrm{E}_1: & a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\
\mathrm{E}_2: & a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\
 & \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\
\mathrm{E}_n: & a_{n1}x_1 + \cdots + a_{nn}x_n = b_n
\end{array}
$$

where the **coefficients** a_{jk} and the b_j are given numbers. The system is called **homogeneous** if all the b_j are zero; otherwise it is called **nonhomogeneous.** Using matrix multiplication (Sec. 7.2), we can write (1) as a single vector equation

$$
(2) \qquad \mathbf{A}\mathbf{x} = \mathbf{b}
$$

where the **coefficient matrix** $\mathbf{A} = [a_{jk}]$ is the $n \times n$ matrix

$$
\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad \text{and} \quad \mathbf{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}
$$

are column vectors. The following matrix $\widetilde{\mathbf{A}}$ is called the **augmented matrix** of the system (1):

$$
\widetilde{\mathbf{A}} = [\mathbf{A} \quad \mathbf{b}] = \begin{bmatrix} a_{11} & \cdots & a_{1n} & b_1 \\ a_{21} & \cdots & a_{2n} & b_2 \\ \cdot & \cdots & \cdot & \cdot \\ a_{n1} & \cdots & a_{nn} & b_n \end{bmatrix}.
$$

A **solution** of (1) is a set of numbers $x_1, \cdots, x_n$ that satisfy all the n equations, and a **solution vector** of (1) is a vector $\mathbf{x}$ whose components constitute a solution of (1).

The method of solving such a system by determinants (Cramer's rule in Sec. 7.7) is not practical, even with efficient methods for evaluating the determinants.

A practical method for the solution of a linear system is the so-called *Gauss elimination*, which we shall now discuss (***proceeding independently of Sec. 7.3***).

Gauss Elimination

This standard method for solving linear systems (1) is a systematic process of elimination that reduces (1) to **triangular form** because the system can then be easily solved by **back substitution**. For instance, a triangular system is

$$\begin{aligned} 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7 \\ 6x_3 &= 3 \end{aligned}$$

and back substitution gives $x_3 = \frac{3}{6} = \frac{1}{2}$ from the third equation, then

$$x_2 = \tfrac{1}{8}(-7 - 2x_3) = -1$$

from the second equation, and finally from the first equation

$$x_1 = \tfrac{1}{3}(8 - 5x_2 - 2x_3) = 4.$$

How do we reduce a given system (1) to triangular form? In the first step we ***eliminate*** x_1 from equations E_2 to E_n in (1). We do this by adding (or subtracting) suitable multiples of E_1 to (from) equations $E_2, \cdots, E_n$ and taking the resulting equations, call them $E_2^*, \cdots, E_n^*$ as the new equations. The first equation, E_1, is called the **pivot equation** in this step, and a_{11} is called the **pivot**. This equation is left unaltered. In the second step we take the new second equation E_2^* (which no longer contains x_1) as the pivot equation and use it to ***eliminate*** $\boldsymbol{x_2}$ from E_3^* to E_n^*. And so on. After $n - 1$ steps this gives a triangular system that can be solved by back substitution as just shown. In this way we obtain precisely all solutions of the *given* system (as proved in Sec. 7.3).

The pivot a_{kk} (in step k) ***must be*** different from zero and ***should be*** large in absolute value to avoid roundoff magnification by the multiplication in the elimination. For this we choose as our pivot equation one that has the absolutely largest a_{jk} in column k on or below the main diagonal (actually, the uppermost if there are several such equations). This popular method is called **partial pivoting**. It is used in CASs (e.g., in Maple).

Partial pivoting distinguishes it from **total pivoting**, which involves both row and column interchanges but is hardly used in practice.

Let us illustrate this method with a simple example.

EXAMPLE 1 Gauss Elimination. Partial Pivoting

Solve the system

$$\begin{aligned} E_1: &\qquad\qquad\;\; 8x_2 + 2x_3 = -7 \\ E_2: &\quad 3x_1 + 5x_2 + 2x_3 = 8 \\ E_3: &\quad 6x_1 + 2x_2 + 8x_3 = 26. \end{aligned}$$

Solution. We must pivot since E_1 has no x_1-term. In Column 1, equation E_3 has the largest coefficient. Hence we interchange E_1 and E_3,

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 3x_1 + 5x_2 + 2x_3 &= 8 \\ 8x_2 + 2x_3 &= -7. \end{aligned}$$

Step 1. Elimination of x_1

It would suffice to show the augmented matrix and operate on it. We show both the equations and the augmented matrix. In the first step, the first equation is the pivot equation. Thus

$$\begin{aligned} \text{Pivot 6} \longrightarrow \quad & \boxed{6x_1} + 2x_2 + 8x_3 = 26 \\ \text{Eliminate} \longrightarrow \quad & \boxed{3x_1} + 5x_2 + 2x_3 = 8 \\ & \qquad\;\; 8x_2 + 2x_3 = -7 \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 3 & 5 & 2 & 8 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

To eliminate x_1 from the other equations (here, from the second equation), do:

Subtract $\frac{3}{6} = \frac{1}{2}$ times the pivot equation from the second equation.

The result is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 4x_2 - 2x_3 &= -5 \\ 8x_2 + 2x_3 &= -7 \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 4 & -2 & -5 \\ 0 & 8 & 2 & -7 \end{array}\right].$$

Step 2. Elimination of x_2

The largest coefficient in Column 2 is 8. Hence we take the *new* third equation as the pivot equation, interchanging equations 2 and 3,

$$\begin{aligned} & 6x_1 + 2x_2 + 8x_3 = 26 \\ \text{Pivot 8} \longrightarrow \quad & \qquad\; \boxed{8x_2} + 2x_3 = -7 \\ \text{Eliminate} \longrightarrow \quad & \qquad\; \boxed{4x_2} - 2x_3 = -5 \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 4 & -2 & -5 \end{array}\right].$$

To eliminate x_2 from the third equation, do:

Subtract $\frac{1}{2}$ times the pivot equation from the third equation.

The resulting triangular system is shown below. This is the end of the forward elimination. Now comes the back substitution.

Back substitution. ***Determination of*** x_3, x_2, x_1

The triangular system obtained in Step 2 is

$$\begin{aligned} 6x_1 + 2x_2 + 8x_3 &= 26 \\ 8x_2 + 2x_3 &= -7 \\ -3x_3 &= -\tfrac{3}{2} \end{aligned} \qquad \left[\begin{array}{ccc|c} 6 & 2 & 8 & 26 \\ 0 & 8 & 2 & -7 \\ 0 & 0 & -3 & -\tfrac{3}{2} \end{array}\right].$$

From this system, taking the last equation, then the second equation, and finally the first equation, we compute the solution

$$\begin{aligned} x_3 &= \tfrac{1}{2} \\ x_2 &= \tfrac{1}{8}(-7 - 2x_3) = -1 \\ x_1 &= \tfrac{1}{6}(26 - 2x_2 - 8x_3) = 4. \end{aligned}$$

This agrees with the values given above, before the beginning of the example. ■

The general algorithm for the Gauss elimination is shown in Table 20.1. To help explain the algorithm, we have numbered some of its lines. b_j is denoted by $a_{j,n+1}$, for uniformity. In lines 1 and 2 we look for a possible pivot. [For $k = 1$ we can always find one; otherwise x_1 would not occur in (1).] In line 2 we do pivoting if necessary, picking an a_{jk} of greatest absolute value (the one with the smallest j if there are several) and interchange the

corresponding rows. If $|a_{kk}|$ is greatest, we do no pivoting. m_{jk} in line 4 suggests *multiplier*, since these are the factors by which we have to multiply the pivot equation E_k^* in Step k before subtracting it from an equation E_j^* below E_k^* from which we want to eliminate x_k. Here we have written E_k^* and E_j^* to indicate that after Step 1 these are no longer the equations given in (1), but these underwent a change in each step, as indicated in line 5. Accordingly, a_{jk} etc. in all lines refer to the most recent equations, and $j \geqq k$ in line 1 indicates that we leave untouched all the equations that have served as pivot equations in previous steps. For $p = k$ in line 5 we get 0 on the right, as it should be in the elimination,

$$a_{jk} - m_{jk}a_{kk} = a_{jk} - \frac{a_{jk}}{a_{kk}} a_{kk} = 0.$$

In line 3, if the last equation in the *triangular* system is $0 = b_n^* \neq 0$, we have no solution. If it is $0 = b_n^* = 0$, we have no unique solution because we then have fewer equations than unknowns.

EXAMPLE 2 Gauss Elimination in Table 20.1, Sample Computation

In Example 1 we had $a_{11} = 0$, so that pivoting was necessary. The greatest coefficient in Column 1 was a_{31}. Thus $\tilde{j} = 3$ in line 2, and we interchanged E_1 and E_3. Then in lines 4 and 5 we computed $m_{21} = \frac{3}{6} = \frac{1}{2}$ and

$$a_{22} = 5 - \tfrac{1}{2} \cdot 2 = 4, \qquad a_{23} = 2 - \tfrac{1}{2} \cdot 8 = -2, \qquad a_{24} = 8 - \tfrac{1}{2} \cdot 26 = -5,$$

and then $m_{31} = \frac{0}{6} = 0$, so that the third equation $8x_2 + 2x_3 = -7$ did not change in Step 1. In Step 2 ($k = 2$) we had 8 as the greatest coefficient in Column 2, hence $\tilde{j} = 3$. We interchanged equations 2 and 3, computed $m_{32} = -\frac{4}{8} = -\frac{1}{2}$ in line 5, and the $a_{33} = -2 - \frac{1}{2} \cdot 2 = -3$, $a_{34} = -5 - \frac{1}{2}(-7) = -\frac{3}{2}$. This produced the triangular form used in the back substitution. ■

If $a_{kk} = 0$ in Step k, ***we must pivot***. If $|a_{kk}|$ is small, ***we should pivot*** because of roundoff error magnification that may seriously affect accuracy or even produce nonsensical results.

EXAMPLE 3 Difficulty with Small Pivots

The solution of the system

$$\begin{aligned} 0.0004x_1 + 1.402x_2 &= 1.406 \\ 0.4003x_1 - 1.502x_2 &= 2.501 \end{aligned}$$

is $x_1 = 10$, $x_2 = 1$. We solve this system by the Gauss elimination, using four-digit floating-point arithmetic. (4D is for simplicity. Make an 8D-arithmetic example that shows the same.)

(a) Picking the first of the given equations as the pivot equation, we have to multiply this equation by $m = 0.4003/0.0004 = 1001$ and subtract the result from the second equation, obtaining

$$-1405x_2 = -1404.$$

Hence $x_2 = -1404/(-1405) = 0.9993$, and from the first equation, instead of $x_1 = 10$, we get

$$x_1 = \frac{1}{0.0004}(1.406 - 1.402 \cdot 0.9993) = \frac{0.005}{0.0004} = 12.5.$$

This failure occurs because $|a_{11}|$ is small compared with $|a_{12}|$, so that a small roundoff error in x_2 leads to a large error in x_1.

(b) Picking the second of the given equations as the pivot equation, we have to multiply this equation by $0.0004/0.4003 = 0.0009993$ and subtract the result from the first equation, obtaining

$$1.404x_2 = 1.404.$$

Hence $x_2 = 1$, and from the pivot equation $x_1 = 10$. This success occurs because $|a_{21}|$ is not very small compared to $|a_{22}|$, so that a small roundoff error in x_2 would not lead to a large error in x_1. Indeed, for instance, if we had the value $x_2 = 1.002$, we would still have from the pivot equation the good value $x_1 = (2.501 + 1.505)/0.4003 = 10.01$. ■

Table 20.1 Gauss Elimination

ALGORITHM GAUSS ($\tilde{\mathbf{A}} = [a_{jk}] = [\mathbf{A} \quad \mathbf{b}]$)

This algorithm computes a unique solution $\mathbf{x} = [x_j]$ of the system (1) or indicates that (1) has no unique solution.

INPUT: Augmented $n \times (n + 1)$ matrix $\tilde{\mathbf{A}} = [a_{jk}]$, where $a_{j,n+1} = b_j$

OUTPUT: Solution $\mathbf{x} = [x_j]$ of (1) or message that the system (1) has no unique solution

For $k = 1, \cdots, n - 1$, do:

1 $m = k$

For $j = k + 1, \cdots, n$, do:

If $(|a_{mk}| < |a_{jk}|)$ then $m = j$

End

If $a_{mk} = 0$ then OUTPUT "No unique solution exists"

Stop

[*Procedure completed unsuccessfully*]

2 Else exchange row k and row m

3 If $a_{nn} = 0$ then OUTPUT "No unique solution exists."

Stop

Else

4 For $j = k + 1, \cdots, n$, do:

$$m_{jk} := \frac{a_{jk}}{a_{kk}}$$

5 For $p = k + 1, \cdots, n + 1$, do:

$$a_{jp} := a_{jp} - m_{jk}a_{kp}$$

End

End

End

6 $$x_n = \frac{a_{n,n+1}}{a_{nn}} \qquad \text{[\textit{Start back substitution}]}$$

For $i = n - 1, \cdots, 1$, do:

7 $$x_i = \frac{1}{a_{ii}}\left(a_{i,n+1} - \sum_{j=i+1}^{n} a_{ij}x_j\right)$$

End

OUTPUT $\mathbf{x} = [x_j]$. Stop

End GAUSS

Error estimates for the Gauss elimination are discussed in Ref. [E5] listed in App. 1.

Row scaling means the multiplication of each Row j by a suitable scaling factor s_j. It is done in connection with partial pivoting to get more accurate solutions. Despite much research (see Refs. [E9], [E24] in App. 1) and the proposition of several principles, scaling is still not well understood. As a possibility, one can scale for pivot choice only (not in the calculation, to avoid additional roundoff) and take as first pivot the entry a_{j1} for which $|a_{j1}|/|A_j|$ is largest; here A_j is an entry of largest absolute value in Row j. Similarly in the further steps of the Gauss elimination.

For instance, for the system

$$\begin{aligned} 4.0000x_1 + 14020x_2 &= 14060 \\ 0.4003x_1 - 1.502x_2 &= 2.501 \end{aligned}$$

we might pick 4 as pivot, but dividing the first equation by 10^4 gives the system in Example 3, for which the second equation is a better pivot equation.

Operation Count

Quite generally, important factors in judging the quality of a numeric method are

Amount of storage

Amount of time ($\equiv$ number of operations)

Effect of roundoff error

For the Gauss elimination, the operation count for a full matrix (a matrix with relatively many nonzero entries) is as follows. In Step k we eliminate x_k from $n - k$ equations. This needs $n - k$ divisions in computing the m_{jk} (line 3) and $(n - k)(n - k + 1)$ multiplications and as many subtractions (both in line 4). Since we do $n - 1$ steps, k goes from 1 to $n - 1$ and thus the total number of operations in this forward elimination is

$$\begin{aligned} f(n) &= \sum_{k=1}^{n-1} (n-k) + 2\sum_{k=1}^{n-1} (n-k)(n-k+1) \qquad \text{(write } n - k = s\text{)} \\ &= \sum_{s=1}^{n-1} s + 2\sum_{s=1}^{n-1} s(s+1) = \tfrac{1}{2}(n-1)n + \tfrac{2}{3}(n^2-1)n \approx \tfrac{2}{3}n^3 \end{aligned}$$

where $2n^3/3$ is obtained by dropping lower powers of n. We see that $f(n)$ grows about proportional to n^3. We say that $f(n)$ is of *order* n^3 and write

$$f(n) = O(n^3)$$

where O suggests **order**. The general definition of O is as follows. We write

$$f(n) = O(h(n))$$

if the quotients $|f(n)/h(n)|$ and $|h(n)/f(n)|$ remain bounded (do not trail off to infinity) as $n \to \infty$. In our present case, $h(n) = n^3$ and, indeed, $f(n)/n^3 \to \frac{2}{3}$ because the omitted terms divided by n^3 go to zero as $n \to \infty$.

In the back substitution of x_i we make $n - i$ multiplications and as many subtractions, as well as 1 division. Hence the number of operations in the back substitution is

$$b(n) = 2\sum_{i=1}^{n}(n-i) + n = 2\sum_{s=1}^{n} s + n = n(n+1) + n = n^2 + 2n = O(n^2).$$

We see that it grows more slowly than the number of operations in the forward elimination of the Gauss algorithm, so that it is negligible for large systems because it is smaller by a factor n, approximately. For instance, if an operation takes 10^{-9} sec, then the times needed are:

Algorithm	$n = 1000$	$n = 10000$
Elimination	0.7 sec	11 min
Back substitution	0.001 sec	0.1 sec

PROBLEM SET 20.1

APPLICATIONS of linear systems see Secs. 7.1 and 8.2.

1–3 GEOMETRIC INTERPRETATION

Solve graphically and explain geometrically.

1. $x_1 - 4x_2 = 20.1$
$3x_1 + 5x_2 = 5.9$

2. $-5.00x_1 + 8.40x_2 = 0$
$10.25x_1 - 17.22x_2 = 0$

3. $7.2x_1 - 3.5x_2 = 16.0$
$-14.4x_1 + 7.0x_2 = 31.0$

4–16 GAUSS ELIMINATION

Solve the following linear systems by Gauss elimination, with partial pivoting if necessary (but without scaling). Show the intermediate steps. Check the result by substitution. If no solution or more than one solution exists, give a reason.

4. $6x_1 + x_2 = -3$
$4x_1 - 2x_2 = 6$

5. $2x_1 - 8x_2 = -4$
$3x_1 + x_2 = 7$

6. $25.38x_1 - 15.48x_2 = 30.60$
$-14.10x_1 + 8.60x_2 = -17.00$

7. $-3x_1 + 6x_2 - 9x_3 = -46.725$
$x_1 - 4x_2 + 3x_3 = 19.571$
$2x_1 + 5x_2 - 7x_3 = -20.073$

8. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

9. $6x_2 + 13x_3 = 137.86$
$6x_1 - 8x_3 = -85.88$
$13x_1 - 8x_2 = 178.54$

10. $4x_1 + 4x_2 + 2x_3 = 0$
$3x_1 - x_2 + 2x_3 = 0$
$3x_1 + 7x_2 + x_3 = 0$

11. $3.4x_1 - 6.12x_2 - 2.72x_3 = 0$
$-x_1 + 1.80x_2 + 0.80x_3 = 0$
$2.7x_1 - 4.86x_2 + 2.16x_3 = 0$

12. $5x_1 + 3x_2 + x_3 = 2$
$-4x_2 + 8x_3 = -3$
$10x_1 - 6x_2 + 26x_3 = 0$

13.
$$\begin{aligned} 3x_2 + 5x_3 &= 1.20736 \\ 3x_1 - 4x_2 &= -2.34066 \\ 5x_1 + 6x_3 &= -0.329193 \end{aligned}$$

14.
$$\begin{aligned} -47x_1 + 4x_2 - 7x_3 &= -118 \\ 19x_1 - 3x_2 + 2x_3 &= 43 \\ -15x_1 + 5x_2 &= -25 \end{aligned}$$

15.
$$\begin{aligned} 2.2x_2 + 1.5x_3 - 3.3x_4 &= -9.30 \\ 0.2x_1 + 1.8x_2 + 4.2x_4 &= 9.24 \\ -x_1 - 3.1x_2 + 2.5x_3 &= -8.70 \\ 0.5x_1 - 3.8x_3 + 1.5x_4 &= 11.94 \end{aligned}$$

16.
$$\begin{aligned} 3.2x_1 + 1.6x_2 &= -0.8 \\ 1.6x_1 - 0.8x_2 + 2.4x_3 &= 16.0 \\ 2.4x_2 - 4.8x_3 + 3.6x_4 &= -39.0 \\ 3.6x_3 + 2.4x_4 &= 10.2 \end{aligned}$$

17. CAS EXPERIMENT. Gauss Elimination. Write a program for the Gauss elimination with pivoting. Apply it to Probs. 13–16. Experiment with systems whose coefficient determinant is small in absolute value. Also investigate the performance of your program for larger systems of your choice, including sparse systems.

18. TEAM PROJECT. Linear Systems and Gauss Elimination. (a) Existence and uniqueness. Find a and b such that $ax_1 + x_2 = b, x_1 + x_2 = 3$ has (i) a unique solution, (ii) infinitely many solutions, (iii) no solutions.

(b) Gauss elimination and nonexistence. Apply the Gauss elimination to the following two systems and compare the calculations step by step. Explain why the elimination fails if no solution exists.

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 13 \end{aligned}$$

$$\begin{aligned} x_1 + x_2 + x_3 &= 3 \\ 4x_1 + 2x_2 - x_3 &= 5 \\ 9x_1 + 5x_2 - x_3 &= 12. \end{aligned}$$

(c) Zero determinant. Why may a computer program give you the result that a homogeneous linear system has only the trivial solution although you know its coefficient determinant to be zero?

(d) Pivoting. Solve System (A) (below) by the Gauss elimination first without pivoting. Show that for any fixed machine word length and sufficiently small $\epsilon > 0$ the computer gives $x_2 = 1$ and then $x_1 = 0$. What is the exact solution? Its limit as $\epsilon \to 0$? Then solve the system by the Gauss elimination with pivoting. Compare and comment.

(e) Pivoting. Solve System (B) by the Gauss elimination and three-digit rounding arithmetic, choosing (i) the first equation, (ii) the second equation as pivot equation. (Remember to round to 3S after each operation before doing the next, just as would be done on a computer!) Then use four-digit rounding arithmetic in those two calculations. Compare and comment.

(A)
$$\begin{aligned} \epsilon x_1 + x_2 &= 1 \\ x_1 + x_2 &= 2 \end{aligned}$$

(B)
$$\begin{aligned} 4.03x_1 + 2.16x_2 &= -4.61 \\ 6.21x_1 + 3.35x_2 &= -7.19 \end{aligned}$$

20.2 Linear Systems: LU-Factorization, Matrix Inversion

We continue our discussion of numeric methods for solving linear systems of n equations in n unknowns $x_1, \cdots, x_n$,

$$\mathbf{Ax} = \mathbf{b} \tag{1}$$

where $\mathbf{A} = [a_{jk}]$ is the $n \times n$ given coefficient matrix and $\mathbf{x}^{\mathsf{T}} = [x_1, \cdots, x_n]$ and $\mathbf{b}^{\mathsf{T}} = [b_1, \cdots, b_n]$. We present three related methods that are modifications of the Gauss

elimination, which require fewer arithmetic operations. They are named after Doolittle, Crout, and Cholesky and use the idea of the LU-factorization of **A**, which we explain first.

An **LU-factorization** of a given square matrix **A** is of the form

$$\mathbf{A} = \mathbf{LU} \tag{2}$$

where **L** is ***lower triangular*** and **U** is ***upper triangular***. For example,

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 8 & 5 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 \\ 4 & 1 \end{bmatrix} \begin{bmatrix} 2 & 3 \\ 0 & -7 \end{bmatrix}.$$

It can be proved that for any nonsingular matrix (see Sec. 7.8) the rows can be reordered so that the resulting matrix **A** has an LU-factorization (2) in which **L** turns out to be the matrix of the ***multipliers*** m_{jk} of the Gauss elimination, with main diagonal $1, \cdots, 1$, and **U** is the matrix of the triangular system at the end of the Gauss elimination. (See Ref. [E5], pp. 155–156, listed in App. 1.)

The ***crucial idea*** now is that **L** and **U** in (2) can be computed directly, without solving simultaneous equations (thus, without using the Gauss elimination). As a count shows, this needs about $n^3/3$ operations, about half as many as the Gauss elimination, which needs about $2n^3/3$ (see Sec. 20.1). And once we have (2), we can use it for solving $\mathbf{Ax} = \mathbf{b}$ in two steps, involving only about n^2 operations, simply by noting that $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ may be written

$$\text{(a)} \quad \mathbf{Ly} = \mathbf{b} \qquad \text{where} \qquad \text{(b)} \quad \mathbf{Ux} = \mathbf{y} \tag{3}$$

and solving first (3a) for **y** and then (3b) for **x**. Here we can require that **L** have main diagonal $1, \cdots, 1$ as stated before; then this is called **Doolittle's method**.[1] Both systems (3a) and (3b) are triangular, so we can solve them as in the back substitution for the Gauss elimination.

A similar method, **Crout's method**,[2] is obtained from (2) if **U** (instead of **L**) is required to have main diagonal $1, \cdots, 1$. In either case the factorization (2) is unique.

EXAMPLE 1 Doolittle's Method

Solve the system in Example 1 of Sec. 20.1 by Doolittle's method.

Solution. The decomposition (2) is obtained from

$$\mathbf{A} = [a_{jk}] = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} = \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ m_{21} & 1 & 0 \\ m_{31} & m_{32} & 1 \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & u_{13} \\ 0 & u_{22} & u_{23} \\ 0 & 0 & u_{33} \end{bmatrix}$$

[1]MYRICK H. DOOLITTLE (1830–1913). American mathematician employed by the U.S. Coast and Geodetic Survey Office. His method appeared in *U.S. Coast and Geodetic Survey*, 1878, 115–120.

[2]PRESCOTT DURAND CROUT (1907–1984), American mathematician, professor at MIT, also worked at General Electric.

by determining the m_{jk} and u_{jk}, using matrix multiplication. By going through **A** row by row we get successively

$a_{11} = 3 = 1 \cdot u_{11} = u_{11}$	$a_{12} = 5 = 1 \cdot u_{12} = u_{12}$	$a_{13} = 2 = 1 \cdot u_{13} = u_{13}$
$a_{21} = 0 = m_{21}u_{11}$ $m_{21} = 0$	$a_{22} = 8 = m_{21}u_{12} + u_{22}$ $u_{22} = 8$	$a_{23} = 2 = m_{21}u_{13} + u_{23}$ $u_{23} = 2$
$a_{31} = 6 = m_{31}u_{11}$ $= m_{31} \cdot 3$ $m_{31} = 2$	$a_{32} = 2 = m_{31}u_{12} + m_{32}u_{22}$ $= 2 \cdot 5 + m_{32} \cdot 8$ $m_{32} = -1$	$a_{33} = 8 = m_{31}u_{13} + m_{32}u_{23} + u_{33}$ $= 2 \cdot 2 - 1 \cdot 2 + u_{33}$ $u_{33} = 6$

Thus the factorization (2) is

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 6 & 2 & 8 \end{bmatrix} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix}.$$

We first solve $\mathbf{Ly} = \mathbf{b}$, determining $y_1 = 8$, then $y_2 = -7$, then y_3 from $2y_1 - y_2 + y_3 = 16 + 7 + y_3 = 26$; thus (note the interchange in **b** because of the interchange in **A**!)

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 2 & -1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 26 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}.$$

Then we solve $\mathbf{Ux} = \mathbf{y}$, determining $x_3 = \frac{3}{6}$ then x_2, then x_1, that is,

$$\begin{bmatrix} 3 & 5 & 2 \\ 0 & 8 & 2 \\ 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \\ 3 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 4 \\ -1 \\ \frac{1}{2} \end{bmatrix}.$$

This agrees with the solution in Example 1 of Sec. 20.1. ■

Our formulas in Example 1 suggest that for general n the entries of the matrices $\mathbf{L} = [m_{jk}]$ (with main diagonal $1, \cdots, 1$ and m_{jk} suggesting "multiplier") and $\mathbf{U} = [u_{jk}]$ in the **Doolittle method** are computed from

$$(4) \qquad \begin{aligned} u_{1k} &= a_{1k} & k &= 1, \cdots, n \\ m_{j1} &= \frac{a_{j1}}{u_{11}} & j &= 2, \cdots, n \\ u_{jk} &= a_{jk} - \sum_{s=1}^{j-1} m_{js}u_{sk} & k &= j, \cdots, n; \quad j \geqq 2 \\ m_{jk} &= \frac{1}{u_{kk}}\left(a_{jk} - \sum_{s=1}^{k-1} m_{js}u_{sk}\right) & j &= k+1, \cdots, n; \quad k \geqq 2. \end{aligned}$$

Row Interchanges. Matrices, such as

$$\begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$$

have no LU-factorization (try!). This indicates that for obtaining an LU-factorization, row interchanges of $\mathbf{A}$ (and corresponding interchanges in $\mathbf{b}$) may be necessary.

Cholesky's Method

For a ***symmetric, positive definite*** matrix $\mathbf{A}$ (thus $\mathbf{A} = \mathbf{A}^{\mathsf{T}}$, $\mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} > 0$ for all $\mathbf{x} \neq \mathbf{0}$) we can in (2) even choose $\mathbf{U} = \mathbf{L}^{\mathsf{T}}$, thus $u_{jk} = m_{kj}$ (but cannot impose conditions on the main diagonal entries). For example,

$$\text{(5)} \qquad \mathbf{A} = \begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \mathbf{L}\mathbf{L}^{\mathsf{T}} = \begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix}.$$

The popular method of solving $\mathbf{A}\mathbf{x} = \mathbf{b}$ based on this factorization $\mathbf{A} = \mathbf{L}\mathbf{L}^{\mathsf{T}}$ is called **Cholesky's method**.[3] In terms of the entries of $\mathbf{L} = [l_{jk}]$ the formulas for the factorization are

$$\text{(6)} \qquad \begin{aligned} l_{11} &= \sqrt{a_{11}} & & \\ l_{j1} &= \frac{a_{j1}}{l_{11}} & & j = 2, \cdots, n \\ l_{jj} &= \sqrt{a_{jj} - \sum_{s=1}^{j-1} l_{js}^2} & & j = 2, \cdots, n \\ l_{pj} &= \frac{1}{l_{jj}} \left(a_{pj} - \sum_{s=1}^{j-1} l_{js} l_{ps} \right) & & p = j + 1, \cdots, n; \quad j \geqq 2. \end{aligned}$$

If $\mathbf{A}$ is symmetric but not positive definite, this method could still be applied, but then leads to a *complex* matrix $\mathbf{L}$, so that the method becomes impractical.

EXAMPLE 2 Cholesky's Method

Solve by Cholesky's method:

$$\begin{aligned} 4x_1 + 2x_2 + 14x_3 &= 14 \\ 2x_1 + 17x_2 - 5x_3 &= -101 \\ 14x_1 - 5x_2 + 83x_3 &= 155. \end{aligned}$$

---

[3] ANDRÉ-LOUIS CHOLESKY (1875–1918), French military officer, geodecist, and mathematician. Surveyed Crete and North Africa. Died in World War I. His method was published posthumously in *Bulletin Géodésique* in 1924 but received little attention until JOHN TODD (1911–2007) — Irish-American mathematician, numerical analysist, and early pioneer of computer methods in numerics, professor at Caltech, and close personal friend and collaborator of ERWIN KREYSZIG, see [E20]—taught Cholesky's method in his analysis course at King's College, London, in the 1940s.

Solution. From (6) or from the form of the factorization

$$\begin{bmatrix} 4 & 2 & 14 \\ 2 & 17 & -5 \\ 14 & -5 & 83 \end{bmatrix} = \begin{bmatrix} l_{11} & 0 & 0 \\ l_{21} & l_{22} & 0 \\ l_{31} & l_{32} & l_{33} \end{bmatrix} \begin{bmatrix} l_{11} & l_{21} & l_{31} \\ 0 & l_{22} & l_{32} \\ 0 & 0 & l_{33} \end{bmatrix}$$

we compute, in the given order,

$$l_{11} = \sqrt{a_{11}} = 2 \qquad l_{21} = \frac{a_{21}}{l_{11}} = \frac{2}{2} = 1 \qquad l_{31} = \frac{a_{31}}{l_{11}} = \frac{14}{2} = 7$$

$$l_{22} = \sqrt{a_{22} - l_{21}^2} = \sqrt{17 - 1} = 4$$

$$l_{32} = \frac{1}{l_{23}}(a_{32} - l_{31}l_{21}) = \frac{1}{4}(-5 - 7 \cdot 1) = -3$$

$$l_{33} = \sqrt{a_{33} - l_{31}^2 - l_{32}^2} = \sqrt{83 - 7^2 - (-3)^2} = 5.$$

This agrees with (5). We now have to solve $\mathbf{Ly} = \mathbf{b}$, that is,

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & 4 & 0 \\ 7 & -3 & 5 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \begin{bmatrix} 14 \\ -101 \\ 155 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{y} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}.$$

As the second step, we have to solve $\mathbf{Ux} = \mathbf{L}^{\mathsf{T}}\mathbf{x} = \mathbf{y}$, that is,

$$\begin{bmatrix} 2 & 1 & 7 \\ 0 & 4 & -3 \\ 0 & 0 & 5 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 7 \\ -27 \\ 5 \end{bmatrix}. \qquad \text{Solution} \qquad \mathbf{x} = \begin{bmatrix} 3 \\ -6 \\ 1 \end{bmatrix}.$$

■

THEOREM 1

Stability of the Cholesky Factorization

The Cholesky LL^{T}-factorization is numerically stable (as defined in Sec. 19.1).

PROOF We have $a_{jj} = l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2$ by squaring the third formula in (6) and solving it for a_{jj}. Hence for all l_{jk} (note that $l_{jk} = 0$ for $k > j$) we obtain (the inequality being trivial)

$$l_{jk}^2 \leqq l_{j1}^2 + l_{j2}^2 + \cdots + l_{jj}^2 = a_{jj}.$$

That is, l_{jk}^2 is bounded by an entry of $\mathbf{A}$, which means stability against rounding. ■

Gauss–Jordan Elimination. Matrix Inversion

Another variant of the Gauss elimination is the **Gauss–Jordan elimination**, introduced by W. Jordan in 1920, in which back substitution is avoided by additional computations that reduce the matrix to diagonal form, instead of the triangular form in the Gauss elimination. But this reduction from the Gauss triangular to the diagonal form requires more operations than back substitution does, so that the method is *disadvantageous* for solving systems $\mathbf{Ax} = \mathbf{b}$. But it may be used for matrix inversion, where the situation is as follows.

The **inverse** of a nonsingular square matrix **A** may be determined in principle by solving the n systems

$$\mathbf{A}\mathbf{x} = \mathbf{b}_j \qquad (j = 1, \cdots, n) \tag{7}$$

where $\mathbf{b}_j$ is the jth column of the $n \times n$ unit matrix.

However, it is preferable to produce $\mathbf{A}^{-1}$ by operating on the unit matrix **I** in the same way as the Gauss–Jordan algorithm, reducing **A** to **I**. A typical illustrative example of this method is given in Sec. 7.8.

PROBLEM SET 20.2

1–5 DOOLITTLE'S METHOD

Show the factorization and solve by Doolittle's method.

1. $4x_1 + 5x_2 = 14$
$12x_1 + 14x_2 = 36$

2. $2x_1 + 9x_2 = 82$
$3x_1 - 5x_2 = -62$

3. $5x_1 + 4x_2 + x_3 = 6.8$
$10x_1 + 9x_2 + 4x_3 = 17.6$
$10x_1 + 13x_2 + 15x_3 = 38.4$

4. $2x_1 + x_2 + 2x_3 = 0$
$-2x_1 + 2x_2 + x_3 = 0$
$x_1 + 2x_2 - 2x_3 = 18$

5. $3x_1 + 9x_2 + 6x_3 = 4.6$
$18x_1 + 48x_2 + 39x_3 = 27.2$
$9x_1 - 27x_2 + 42x_3 = 9.0$

6. TEAM PROJECT. Crout's method factorizes $\mathbf{A} = \mathbf{L}\mathbf{U}$, where **L** is lower triangular and **U** is upper triangular with diagonal entries $u_{jj} = 1, j = 1, \cdots, n$.

(a) Formulas. Obtain formulas for Crout's method similar to (4).

(b) Examples. Solve Prob. 5 by Crout's method.

(c) Factor the following matrix by the Doolittle, Crout, and Cholesky methods.

$$\begin{bmatrix} 1 & -4 & 2 \\ -4 & 25 & 4 \\ 2 & 4 & 24 \end{bmatrix}$$

(d) Give the formulas for factoring a tridiagonal matrix by Crout's method.

(e) When can you obtain Crout's factorization from Doolittle's by transposition?

7–12 CHOLESKY'S METHOD

Show the factorization and solve.

7. $9x_1 + 6x_2 + 12x_3 = 17.4$
$6x_1 + 13x_2 + 11x_3 = 23.6$
$12x_1 + 11x_2 + 26x_3 = 30.8$

8. $4x_1 + 6x_2 + 8x_3 = 0$
$6x_1 + 34x_2 + 52x_3 = -160$
$8x_1 + 52x_2 + 129x_3 = -452$

9. $0.01x_1 + 0.03x_3 = 0.14$
$0.16x_2 + 0.08x_3 = 0.16$
$0.03x_1 + 0.08x_2 + 0.14x_3 = 0.54$

10. $4x_1 + 2x_3 = 1.5$
$4x_2 + x_3 = 4.0$
$2x_1 + x_2 + 2x_3 = 2.5$

11. $x_1 - x_2 + 3x_3 + 2x_4 = 15$
$-x_1 + 5x_2 - 5x_3 - 2x_4 = -35$
$3x_1 - 5x_2 + 19x_3 + 3x_4 = 94$
$2x_1 - 2x_2 + 3x_3 + 21x_4 = 1$

12. $4x_1 + 2x_2 + 4x_3 = 20$
$2x_1 + 2x_2 + 3x_3 + 2x_4 = 36$
$4x_1 + 3x_2 + 6x_3 + 3x_4 = 60$
$2x_2 + 3x_3 + 9x_4 = 122$

13. Definiteness. Let **A**, **B** be $n \times n$ and positive definite. Are $-\mathbf{A}$, $\mathbf{A}^\mathsf{T}$, $\mathbf{A} + \mathbf{B}$, $\mathbf{A} - \mathbf{B}$ positive definite?

14. CAS PROJECT. Cholesky's Method. (a) Write a program for solving linear systems by Cholesky's method and apply it to Example 2 in the text, to Probs. 7–9, and to systems of your choice.

(b) Splines. Apply the factorization part of the program to the following matrices (as they occur in (9), Sec. 19.4 (with $c_j = 1$), in connection with splines).

$$\begin{bmatrix} 2 & 1 & 0 \\ 1 & 4 & 1 \\ 0 & 1 & 2 \end{bmatrix}, \quad \begin{bmatrix} 2 & 1 & 0 & 0 \\ 1 & 4 & 1 & 0 \\ 0 & 1 & 4 & 1 \\ 0 & 0 & 1 & 2 \end{bmatrix}.$$

15–19 INVERSE

Find the inverse by the Gauss–Jordan method, showing the details.

15. In Prob. 1

16. In Prob. 4

17. In Team Project 6(c)

18. In Prob. 9

19. In Prob. 12

20. Rounding. For the following matrix $\mathbf{A}$ find det $\mathbf{A}$. What happens if you roundoff the given entries to **(a)** 5S, **(b)** 4S, **(c)** 3S, **(d)** 2S, **(e)** 1S? What is the practical implication of your work?

$$\mathbf{A} = \begin{bmatrix} \frac{1}{3} & \frac{1}{4} & 2 \\ -\frac{1}{9} & 1 & \frac{1}{7} \\ \frac{4}{63} & -\frac{3}{28} & \frac{13}{49} \end{bmatrix}$$

20.3 Linear Systems: Solution by Iteration

The Gauss elimination and its variants in the last two sections belong to the **direct methods** for solving linear systems of equations; these are methods that give solutions after an amount of computation that can be specified in advance. In contrast, in an **indirect** or **iterative method** we start from an approximation to the true solution and, if successful, obtain better and better approximations from a computational cycle repeated as often as may be necessary for achieving a required accuracy, so that the amount of arithmetic depends upon the accuracy required and varies from case to case.

We apply iterative methods if the convergence is rapid (if matrices have large main diagonal entries, as we shall see), so that we save operations compared to a direct method. We also use iterative methods if a large system is **sparse**, that is, has very many zero coefficients, so that one would waste space in storing zeros, for instance, 9995 zeros per equation in a potential problem of 10^4 equations in 10^4 unknowns with typically only 5 nonzero terms per equation (more on this in Sec. 21.4).

Gauss–Seidel Iteration Method[4]

This is an iterative method of great practical importance, which we can simply explain in terms of an example.

EXAMPLE 1 Gauss–Seidel Iteration

We consider the linear system

$$\begin{aligned} x_1 - 0.25x_2 - 0.25x_3 \qquad\qquad &= 50 \\ -0.25x_1 + x_2 \qquad\qquad - 0.25x_4 &= 50 \\ -0.25x_1 \qquad\qquad + x_3 - 0.25x_4 &= 25 \\ - 0.25x_2 - 0.25x_3 + x_4 &= 25. \end{aligned} \tag{1}$$

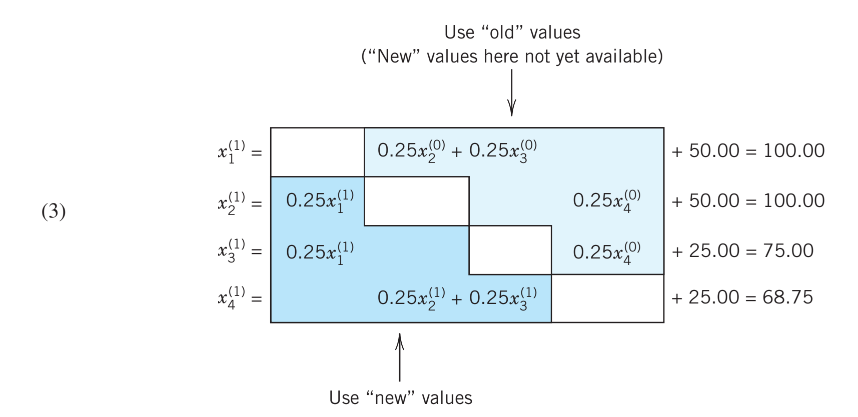
[4]PHILIPP LUDWIG VON SEIDEL (1821–1896), German mathematician. For Gauss see footnote 5 in Sec. 5.4.

(Equations of this form arise in the numeric solution of PDEs and in spline interpolation.) We write the system in the form

$$
\begin{aligned}
x_1 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 50 \\
x_2 &= 0.25x_1 \phantom{{}+ 0.25x_2 + 0.25x_3} + 0.25x_4 + 50 \\
x_3 &= 0.25x_1 \phantom{{}+ 0.25x_2 + 0.25x_3} + 0.25x_4 + 25 \\
x_4 &= \phantom{0.25x_1 +{}} 0.25x_2 + 0.25x_3 \phantom{{}+ 0.25x_4} + 25.
\end{aligned}
\tag{2}
$$

These equations are now used for iteration; that is, we start from a (possibly poor) approximation to the solution, say $x_1^{(0)} = 100, x_2^{(0)} = 100, x_3^{(0)} = 100, x_4^{(0)} = 100$, and compute from (2) a perhaps better approximation

(3)



These equations (3) are obtained from (2) by substituting on the right the ***most recent*** approximation for each unknown. In fact, corresponding values replace previous ones as soon as they have been computed, so that in the second and third equations we use $x_1^{(1)}$ (not $x_1^{(0)}$), and in the last equation of (3) we use $x_2^{(1)}$ and $x_3^{(1)}$ (not $x_2^{(0)}$ and $x_3^{(0)}$). Using the same principle, we obtain in the next step

$$
\begin{aligned}
x_1^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(1)} + 0.25x_3^{(1)} \phantom{{}+ 0.25x_4^{(1)}} + 50.00 = 93.750 \\
x_2^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(1)} + 0.25x_3^{(1)}} + 0.25x_4^{(1)} + 50.00 = 90.625 \\
x_3^{(2)} &= 0.25x_1^{(2)} \phantom{{}+ 0.25x_2^{(1)} + 0.25x_3^{(1)}} + 0.25x_4^{(1)} + 25.00 = 65.625 \\
x_4^{(2)} &= \phantom{0.25x_1^{(2)} +{}} 0.25x_2^{(2)} + 0.25x_3^{(2)} \phantom{{}+ 0.25x_4^{(1)}} + 25.00 = 64.062
\end{aligned}
$$

Further steps give the values

x_1	x_2	x_3	x_4
89.062	88.281	63.281	62.891
87.891	87.695	62.695	62.598
87.598	87.549	62.549	62.524
87.524	87.512	62.512	62.506
87.506	87.503	62.503	62.502

Hence convergence to the exact solution $x_1 = x_2 = 87.5, x_3 = x_4 = 62.5$ (verify!) seems rather fast. ■

An algorithm for the Gauss–Seidel iteration is shown in Table 20.2. To obtain the algorithm, let us derive the general formulas for this iteration.

We assume that $a_{jj} = 1$ for $j = 1, \cdots, n$. (Note that this can be achieved if we can rearrange the equations so that no diagonal coefficient is zero; then we may divide each equation by the corresponding diagonal coefficient.) We now write

(4) $$\mathbf{A} = \mathbf{I} + \mathbf{L} + \mathbf{U} \qquad (a_{jj} = 1)$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{L}$ and $\mathbf{U}$ are, respectively, lower and upper triangular matrices with zero main diagonals. If we substitute (4) into $\mathbf{Ax} = \mathbf{b}$, we have

$$\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}.$$

Taking $\mathbf{Lx}$ and $\mathbf{Ux}$ to the right, we obtain, since $\mathbf{Ix} = \mathbf{x}$,

(5) $$\mathbf{x} = \mathbf{b} - \mathbf{Lx} - \mathbf{Ux}.$$

Remembering from (3) in Example 1 that below the main diagonal we took "new" approximations and above the main diagonal "old" ones, we obtain from (5) the desired iteration formulas

"New" "Old"

(6) $$\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{Lx}^{(m+1)} - \mathbf{Ux}^{(m)} \qquad (a_{jj} = 1)$$

where $\mathbf{x}^{(m)} = [x_j^{(m)}]$ is the mth approximation and $\mathbf{x}^{(m+1)} = [x_j^{(m+1)}]$ is the $(m + 1)$st approximation. In components this gives the formula in line 1 in Table 20.2. The matrix $\mathbf{A}$ must satisfy $a_{jj} \neq 0$ for all j. In Table 20.2 our assumption $a_{jj} = 1$ is no longer required, but is automatically taken care of by the factor $1/a_{jj}$ in line 1.

Table 20.2 Gauss–Seidel Iteration

ALGORITHM GAUSS–SEIDEL ($\mathbf{A}, \mathbf{b}, \mathbf{x}^{(0)}, \epsilon, N$)

This algorithm computes a solution $\mathbf{x}$ of the system $\mathbf{Ax} = \mathbf{b}$ given an initial approximation $\mathbf{x}^{(0)}$, where $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix with $a_{jj} \neq 0, j = 1, \cdots, n$.

INPUT: $\mathbf{A}, \mathbf{b}$, initial approximation $\mathbf{x}^{(0)}$, tolerance $\epsilon > 0$, maximum number of iterations N

OUTPUT: Approximate solution $\mathbf{x}^{(m)} = [x_j^{(m)}]$ or failure message that $\mathbf{x}^{(N)}$ does not satisfy the tolerance condition

For $m = 0, \cdots, N - 1$, do:

For $j = 1, \cdots, n$, do:

1 $$x_j^{(m+1)} = \frac{1}{a_{jj}}\left(b_j - \sum_{k=1}^{j-1} a_{jk}x_k^{(m+1)} - \sum_{k=j+1}^{n} a_{jk}x_k^{(m)}\right)$$

End

2 If $\max_j |x_j^{(m+1)} - x_j^{(m)}| < \epsilon\, |x_j^{(m+1)}|$ then OUTPUT $\mathbf{x}^{(m+1)}$. Stop

[*Procedure completed successfully*]

End

OUTPUT: "No solution satisfying the tolerance condition obtained after N iteration steps." Stop

[*Procedure completed unsuccessfully*]

End GAUSS–SEIDEL

Convergence and Matrix Norms

An iteration method for solving $\mathbf{Ax} = \mathbf{b}$ is said to **converge** for an initial $\mathbf{x}^{(0)}$ if the corresponding iterative sequence $\mathbf{x}^{(0)}, \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \cdots$ converges to a solution of the given system. Convergence depends on the relation between $\mathbf{x}^{(m)}$ and $\mathbf{x}^{(m+1)}$. To get this relation for the Gauss–Seidel method, we use (6). We first have

$$(\mathbf{I} + \mathbf{L})\,\mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{U}\mathbf{x}^{(m)}$$

and by multiplying by $(\mathbf{I} + \mathbf{L})^{-1}$ from the left,

$$\mathbf{x}^{(m+1)} = \mathbf{C}\mathbf{x}^{(m)} + (\mathbf{I} + \mathbf{L})^{-1}\mathbf{b} \qquad \text{where} \qquad \mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}. \tag{7}$$

The Gauss–Seidel iteration converges for every $\mathbf{x}^{(0)}$ if and only if all the eigenvalues (Sec. 8.1) of the "iteration matrix" $\mathbf{C} = [c_{jk}]$ have absolute value less than 1. (Proof in Ref. [E5], p. 191, listed in App. 1.)

CAUTION! If you want to get $\mathbf{C}$, first divide the rows of $\mathbf{A}$ by a_{jj} to have main diagonal $1, \cdots, 1$. If the **spectral radius** of $\mathbf{C}$ (= maximum of those absolute values) is small, then the convergence is rapid.

Sufficient Convergence Condition. A sufficient condition for convergence is

$$\|\mathbf{C}\| < 1. \tag{8}$$

Here $\|\mathbf{C}\|$ is some **matrix norm**, such as

$$\|\mathbf{C}\| = \sqrt{\sum_{j=1}^{n} \sum_{k=1}^{n} c_{jk}^2} \qquad \textbf{(Frobenius norm)} \tag{9}$$

or the greatest of the sums of the $|c_{jk}|$ in a *column* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_k \sum_{j=1}^{n} |c_{jk}| \qquad \textbf{(Column "sum" norm)} \tag{10}$$

or the greatest of the sums of the $|c_{jk}|$ in a *row* of $\mathbf{C}$

$$\|\mathbf{C}\| = \max_j \sum_{k=1}^{n} |c_{jk}| \qquad \textbf{(Row "sum" norm).} \tag{11}$$

These are the most frequently used matrix norms in numerics.

In most cases the choice of one of these norms is a matter of computational convenience. However, the following example shows that sometimes one of these norms is preferable to the others.

EXAMPLE 2 Test of Convergence of the Gauss–Seidel Iteration

Test whether the Gauss–Seidel iteration converges for the system

$$\begin{aligned} 2x + y + z &= 4 \\ x + 2y + z &= 4 \\ x + y + 2z &= 4 \end{aligned} \qquad \text{written} \qquad \begin{aligned} x &= 2 - \tfrac{1}{2}y - \tfrac{1}{2}z \\ y &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}z \\ z &= 2 - \tfrac{1}{2}x - \tfrac{1}{2}y. \end{aligned}$$

Solution. The decomposition (multiply the matrix by $\frac{1}{2}$ – why?) is

$$\begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 1 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 \end{bmatrix} = \mathbf{I} + \mathbf{L} + \mathbf{U} = \mathbf{I} + \begin{bmatrix} 0 & 0 & 0 \\ \frac{1}{2} & 0 & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix} + \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}.$$

It shows that

$$\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U} = -\begin{bmatrix} 1 & 0 & 0 \\ -\frac{1}{2} & 1 & 0 \\ -\frac{1}{4} & -\frac{1}{2} & 1 \end{bmatrix} \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -\frac{1}{2} & -\frac{1}{2} \\ 0 & \frac{1}{4} & -\frac{1}{4} \\ 0 & \frac{1}{8} & \frac{3}{8} \end{bmatrix}.$$

We compute the Frobenius norm of $\mathbf{C}$

$$\|\mathbf{C}\| = \left(\tfrac{1}{4} + \tfrac{1}{4} + \tfrac{1}{16} + \tfrac{1}{16} + \tfrac{1}{64} + \tfrac{9}{64}\right)^{1/2} = \left(\tfrac{50}{64}\right)^{1/2} = 0.884 < 1$$

and conclude from (8) that this Gauss–Seidel iteration converges. It is interesting that the other two norms would permit no conclusion, as you should verify. Of course, this points to the fact that (8) is sufficient for convergence rather than necessary. ■

Residual. Given a system $\mathbf{Ax} = \mathbf{b}$, the **residual** $\mathbf{r}$ of $\mathbf{x}$ with respect to this system is defined by

$$\mathbf{r} = \mathbf{b} - \mathbf{Ax}. \qquad (12)$$

Clearly, $\mathbf{r} = \mathbf{0}$ if and only if $\mathbf{x}$ is a solution. Hence $\mathbf{r} \neq \mathbf{0}$ for an approximate solution. In the Gauss–Seidel iteration, at each stage we modify or *relax* a component of an approximate solution in order to reduce a component of $\mathbf{r}$ to zero. Hence the Gauss–Seidel iteration belongs to a class of methods often called **relaxation methods**. More about the residual follows in the next section.

Jacobi Iteration

The Gauss–Seidel iteration is a method of **successive corrections** because for each component we successively replace an approximation of a component by a corresponding new approximation as soon as the latter has been computed. An iteration method is called a method of **simultaneous corrections** if no component of an approximation $\mathbf{x}^{(m)}$ is used until ***all*** the components of $\mathbf{x}^{(m)}$ have been computed. A method of this type is the **Jacobi iteration**, which is similar to the Gauss–Seidel iteration but involves *not* using improved values until a step has been completed and then replacing $\mathbf{x}^{(m)}$ by $\mathbf{x}^{(m+1)}$ at once, directly before the beginning of the next step. Hence if we write $\mathbf{Ax} = \mathbf{b}$ (***with*** $a_{jj} = 1$ ***as before!***) in the form $\mathbf{x} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}$, the Jacobi iteration in matrix notation is

$$\mathbf{x}^{(m+1)} = \mathbf{b} + (\mathbf{I} - \mathbf{A})\mathbf{x}^{(m)} \qquad (a_{jj} = 1). \qquad (13)$$

This method converges for every choice of $\mathbf{x}^{(0)}$ if and only if the spectral radius of $\mathbf{I} - \mathbf{A}$ is less than 1. It has recently gained greater practical interest since on parallel processors all n equations can be solved simultaneously at each iteration step.

For Jacobi, see Sec. 10.3. For exercises, see the problem set.

PROBLEM SET 20.3

1. Verify the solution in Example 1 of the text.

2. Show that for the system in Example 2 the Jacobi iteration diverges. *Hint.* Use eigenvalues.

3. Verify the claim at the end of Example 2.

4–10 GAUSS–SEIDEL ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ and using 6S in the computation. *Hint.* Make sure that you solve each equation for the variable that has the largest coefficient (why?). Show the details.

4.
$$\begin{aligned} 4x_1 - x_2 \qquad &= 21 \\ -x_1 + 4x_2 - x_3 &= -45 \\ - x_2 + 4x_3 &= 33 \end{aligned}$$

5.
$$\begin{aligned} 10x_1 + x_2 + x_3 &= 6 \\ x_1 + 10x_2 + x_3 &= 6 \\ x_1 + x_2 + 10x_3 &= 6 \end{aligned}$$

6.
$$\begin{aligned} x_2 + 7x_3 &= 25.5 \\ 5x_1 + x_2 \qquad &= 0 \\ x_1 + 6x_2 + x_3 &= -10.5 \end{aligned}$$

7.
$$\begin{aligned} 5x_1 - 2x_2 \qquad &= 18 \\ -2x_1 + 10x_2 - 2x_3 &= -60 \\ - 2x_2 + 15x_3 &= 128 \end{aligned}$$

8.
$$\begin{aligned} 3x_1 + 2x_2 + x_3 &= 7 \\ x_1 + 3x_2 + 2x_3 &= 4 \\ 2x_1 + x_2 + 3x_3 &= 7 \end{aligned}$$

9.
$$\begin{aligned} 5x_1 + x_2 + 2x_3 &= 19 \\ x_1 + 4x_2 - 2x_3 &= -2 \\ 2x_1 + 3x_2 + 8x_3 &= 39 \end{aligned}$$

10.
$$\begin{aligned} 4x_1 \qquad + 5x_3 &= 12.5 \\ x_1 + 6x_2 + 2x_3 &= 18.5 \\ 8x_1 + 2x_2 + x_3 &= -11.5 \end{aligned}$$

11. Apply the Gauss–Seidel iteration (3 steps) to the system in Prob. 5, starting from **(a)** 0, 0, 0 **(b)** 10, 10, 10. Compare and comment.

12. In Prob. 5, compute $\mathbf{C}$ **(a)** if you solve the first equation for x_1, the second for x_2, the third for x_3, proving convergence; **(b)** if you nonsensically solve the third equation for x_1, the first for x_2, the second for x_3, proving divergence.

13. CAS Experiment. Gauss–Seidel Iteration. (a) Write a program for Gauss–Seidel iteration.

(b) Apply the program $\mathbf{A}(t)\mathbf{x} = \mathbf{b}$, to starting from $[0 \quad 0 \quad 0]^\mathsf{T}$, where

$$\mathbf{A}(t) = \begin{bmatrix} 1 & t & t \\ t & 1 & t \\ t & t & 1 \end{bmatrix}, \qquad \mathbf{b} = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}.$$

For $t = 0.2, 0.5, 0.8, 0.9$ determine the number of steps to obtain the exact solution to 6S and the corresponding spectral radius of $\mathbf{C}$. Graph the number of steps and the spectral radius as functions of t and comment.

(c) Successive overrelaxation (SOR). Show that by adding and subtracting $\mathbf{x}^{(m)}$ on the right, formula (6) can be written

$$\mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)} \qquad (a_{jj} = 1).$$

Anticipation of further corrections motivates the introduction of an **overrelaxation factor** $\omega > 1$ to get the **SOR formula for Gauss–Seidel**

$$\text{(14)} \qquad \begin{aligned} \mathbf{x}^{(m+1)} = \mathbf{x}^{(m)} + \omega(\mathbf{b} &- \mathbf{L}\mathbf{x}^{(m+1)} \\ &- (\mathbf{U} + \mathbf{I})\mathbf{x}^{(m)}) \qquad (a_{jj} = 1) \end{aligned}$$

intended to give more rapid convergence. A recommended value is $\omega = 2/(1 + \sqrt{1 - \rho})$, where ρ is the spectral radius of $\mathbf{C}$ in (7). Apply SOR to the matrix in (b) for $t = 0.5$ and 0.8 and notice the improvement of convergence. (Spectacular gains are made with larger systems.)

14–17 JACOBI ITERATION

Do 5 steps, starting from $\mathbf{x}_0 = [1 \quad 1 \quad 1]$. Compare with the Gauss–Seidel iteration. Which of the two seems to converge faster? Show the details of your work.

14. The system in Prob. 4

15. The system in Prob. 9

16. The system in Prob. 10

17. Show convergence in Prob. 16 by verifying that $\mathbf{I} - \mathbf{A}$, where $\mathbf{A}$ is the matrix in Prob. 16 with the rows divided by the corresponding main diagonal entries, has the eigenvalues -0.519589 and $0.259795 \pm 0.246603i$.

18–20 NORMS

Compute the norms (9), (10), (11) for the following (square) matrices. Comment on the reasons for greater or smaller differences among the three numbers.

18. The matrix in Prob. 10

19. The matrix in Prob. 5

20. $\begin{bmatrix} 2k & -k & -k \\ k & -2k & k \\ -k & -k & 2k \end{bmatrix}$

20.4 Linear Systems: Ill-Conditioning, Norms

One does not need much experience to observe that some systems $\mathbf{Ax} = \mathbf{b}$ are good, giving accurate solutions even under roundoff or coefficient inaccuracies, whereas others are bad, so that these inaccuracies affect the solution strongly. We want to see what is going on and whether or not we can "trust" a linear system. Let us first formulate the two relevant concepts (ill- and well-conditioned) for general numeric work and then turn to linear systems and matrices.

A computational problem is called **ill-conditioned** (or *ill-posed*) if "small" changes in the data (the input) cause "large" changes in the solution (the output). On the other hand, a problem is called **well-conditioned** (or *well-posed*) if "small" changes in the data cause only "small" changes in the solution.

These concepts are qualitative. We would certainly regard a magnification of inaccuracies by a factor 100 as "large," but could debate where to draw the line between "large" and "small," depending on the kind of problem and on our viewpoint. Double precision may sometimes help, but if data are measured inaccurately, one should attempt ***changing the mathematical setting*** of the problem to a well-conditioned one.

Let us now turn to linear systems. Figure 445 explains that ill-conditioning occurs if and only if the two equations give two nearly parallel lines, so that their intersection point (the solution of the system) moves substantially if we raise or lower a line just a little. For larger systems the situation is similar in principle, although geometry no longer helps. We shall see that we may regard ill-conditioning as an approach to singularity of the matrix.

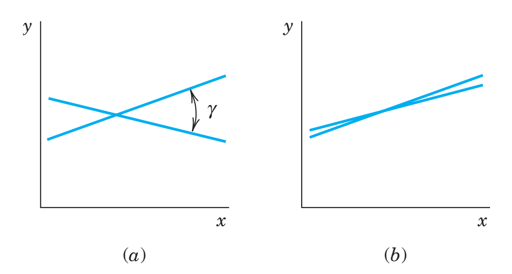


Fig. 445. (a) Well-conditioned and (b) ill-conditioned linear system of two equations in two unknowns

EXAMPLE 1 An Ill-Conditioned System

You may verify that the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \quad y &= 1 \end{aligned}$$

has the solution $x = 0.5$, $y = -0.5$, whereas the system

$$\begin{aligned} 0.9999x - 1.0001y &= 1 \\ x - \quad y &= 1 + \epsilon \end{aligned}$$

has the solution $x = 0.5 + 5000.5\epsilon$, $y = -0.5 + 4999.5\epsilon$. This shows that the system is ill-conditioned because a change on the right of magnitude ϵ produces a change in the solution of magnitude 5000ϵ, approximately. We see that the lines given by the equations have nearly the same slope. ■

Well-conditioning can be asserted if the main diagonal entries of $\mathbf{A}$ have large absolute values compared to those of the other entries. Similarly if $\mathbf{A}^{-1}$ and $\mathbf{A}$ have maximum entries of about the same absolute value.

Ill-conditioning is indicated if $\mathbf{A}^{-1}$ has entries of large absolute value compared to those of the solution (about 5000 in Example 1) and if poor approximate solutions may still produce small residuals.

Residual. The *residual* $\mathbf{r}$ of an approximate solution $\tilde{\mathbf{x}}$ of $\mathbf{Ax} = \mathbf{b}$ is defined as

$$\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}. \tag{1}$$

Now $\mathbf{b} = \mathbf{Ax}$, so that

$$\mathbf{r} = \mathbf{A}(\mathbf{x} - \mathbf{A}\tilde{\mathbf{x}}). \tag{2}$$

Hence $\mathbf{r}$ is small if $\tilde{\mathbf{x}}$ has high accuracy, but the converse may be false:

EXAMPLE 2 Inaccurate Approximate Solution with a Small Residual

The system

$$\begin{aligned} 1.0001x_1 + \quad x_2 &= 2.0001 \\ x_1 + 1.0001x_2 &= 2.0001 \end{aligned}$$

has the exact solution $x_1 = 1, x_2 = 1$. Can you see this by inspection? The very inaccurate approximation $\tilde{x}_1 = 2.0000, \tilde{x}_2 = 0.0001$ has the very small residual (to 4D)

$$\mathbf{r} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 1.0001 & 1.0000 \\ 1.0000 & 1.0001 \end{bmatrix} \begin{bmatrix} 2.0000 \\ 0.0001 \end{bmatrix} = \begin{bmatrix} 2.0001 \\ 2.0001 \end{bmatrix} - \begin{bmatrix} 2.0003 \\ 2.0001 \end{bmatrix} = \begin{bmatrix} -0.0002 \\ 0.0000 \end{bmatrix}.$$

From this, a naive person might draw the false conclusion that the approximation should be accurate to 3 or 4 decimals.

Our result is probably unexpected, but we shall see that it has to do with the fact that the system is ill-conditioned. ■

Our goal is to show that ill-conditioning of a linear system and of its coefficient matrix $\mathbf{A}$ can be measured by a number, the *condition number* $\kappa(\mathbf{A})$. Other measures for ill-conditioning

have also been proposed, but $\kappa(\mathbf{A})$ is probably the most widely used one. $\kappa(\mathbf{A})$ is defined in terms of norm, a concept of great general interest throughout numerics (and in modern mathematics in general!). We shall reach our goal in three steps, discussing

1. **Vector norms**
2. **Matrix norms**
3. **Condition number** κ of a square matrix

Vector Norms

A **vector norm** for column vectors $\mathbf{x} = [x_j]$ with n components (n fixed) is a generalized length or distance. It is denoted by $\|\mathbf{x}\|$ and is defined by four properties of the usual length of vectors in three-dimensional space, namely,

$$\text{(3)}\quad \begin{aligned} &\text{(a)} && \|\mathbf{x}\| \text{ is a nonnegative real number.} \\ &\text{(b)} && \|\mathbf{x}\| = 0 \quad \text{if and only if} \quad \mathbf{x} = \mathbf{0}. \\ &\text{(c)} && \|k\mathbf{x}\| = |k|\,\|\mathbf{x}\| \qquad \text{for all } k. \\ &\text{(d)} && \|\mathbf{x} + \mathbf{y}\| \leqq \|\mathbf{x}\| + \|\mathbf{y}\| \qquad \text{(Triangle inequality).} \end{aligned}$$

If we use several norms, we label them by a subscript. Most important in connection with computations is the ***p-norm*** defined by

$$\text{(4)}\qquad \|\mathbf{x}\|_p = (|x_1|^p + |x_2|^p + \cdots + |x_n|^p)^{1/p}$$

where p is a fixed number and $p \geqq 1$. In practice, one usually takes $p = 1$ or 2 and, as a third norm, $\|\mathbf{x}\|_\infty$ (the latter as defined below), that is,

$$\textbf{(5)}\qquad \|\mathbf{x}\|_1 = |x_1| + \cdots + |x_n| \qquad (\text{“}l_1\text{-norm”})$$

$$\textbf{(6)}\qquad \|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2} \qquad (\text{“Euclidean” or “}l_2\text{-norm”})$$

$$\textbf{(7)}\qquad \|\mathbf{x}\|_\infty = \max_j |x_j| \qquad (\text{“}l_\infty\text{-norm”}).$$

For $n = 3$ the l_2-norm is the usual length of a vector in three-dimensional space. The l_1-norm and l_∞-norm are generally more convenient in computation. But all three norms are in common use.

EXAMPLE 3 Vector Norms

If $\mathbf{x}^\mathsf{T} = [2 \quad -3 \quad 0 \quad 1 \quad -4]$, then $\|\mathbf{x}\|_1 = 10$, $\|\mathbf{x}\|_2 = \sqrt{30}$, $\|\mathbf{x}\|_\infty = 4$. ■

In three-dimensional space, two points with position vectors $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have distance $|\mathbf{x} - \tilde{\mathbf{x}}|$ from each other. For a linear system $\mathbf{A}\mathbf{x} = \mathbf{b}$, this suggests that we take $\|\mathbf{x} - \tilde{\mathbf{x}}\|$ as a measure of inaccuracy and call it the **distance** between an exact and an approximate solution, or the **error** of $\tilde{\mathbf{x}}$.

Matrix Norm

If $\mathbf{A}$ is an $n \times n$ matrix and $\mathbf{x}$ any vector with n components, then $\mathbf{A}\mathbf{x}$ is a vector with n components. We now take a vector norm and consider $\|\mathbf{x}\|$ and $\|\mathbf{A}\mathbf{x}\|$. One can prove (see

Ref. [E17]. pp. 77, 92–93, listed in App. 1) that there is a number c (depending on $\mathbf{A}$) such that

$$\|\mathbf{Ax}\| \leqq c\|\mathbf{x}\| \qquad \text{for all } \mathbf{x}. \tag{8}$$

Let $\mathbf{x} \neq 0$. Then $\|\mathbf{x}\| > 0$ by (3b) and division gives $\|\mathbf{Ax}\|/\|\mathbf{x}\| \leqq c$. We obtain the smallest possible c valid for *all* $\mathbf{x}$ ($\neq \mathbf{0}$) by taking the maximum on the left. This smallest c is called the **matrix norm of A** *corresponding to the vector norm we picked* and is denoted by $\|\mathbf{A}\|$. Thus

$$\|\mathbf{A}\| = \max \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \qquad (\mathbf{x} \neq \mathbf{0}), \tag{9}$$

the maximum being taken over all $\mathbf{x} \neq \mathbf{0}$. Alternatively [see (c) in Team Project 24],

$$\|\mathbf{A}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|. \tag{10}$$

The maximum in (10) and thus also in (9) exists. And the name "matrix ***norm***" is justified because $\|\mathbf{A}\|$ satisfies (3) with $\mathbf{x}$ and $\mathbf{y}$ replaced by $\mathbf{A}$ and $\mathbf{B}$. (Proofs in Ref. [E17] pp. 77, 92–93.)

Note carefully that $\|\mathbf{A}\|$ depends on the vector norm that we selected. In particular, one can show that

for the l_1-norm (5) one gets the column "sum" norm (10), Sec. 20.3,

for the l_∞-norm (7) one gets the row "sum" norm (11), Sec. 20.3.

By taking our best possible (our smallest) $c = \|\mathbf{A}\|$ we have from (8)

$$\|\mathbf{Ax}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|. \tag{11}$$

This is the formula we shall need. Formula (9) also implies for two $n \times n$ matrices (see Ref. [E17], p. 98)

$$\|\mathbf{AB}\| \leqq \|\mathbf{A}\|\,\|\mathbf{B}\|, \qquad \text{thus} \qquad \|\mathbf{A}^n\| \leqq \|\mathbf{A}\|^n. \tag{12}$$

See Refs. [E9] and [E17] for other useful formulas on norms.

Before we go on, let us do a simple illustrative computation.

EXAMPLE 4 Matrix Norms

Compute the matrix norms of the coefficient matrix $\mathbf{A}$ in Example 1 and of its inverse $\mathbf{A}^{-1}$, assuming that we use (a) the l_1-vector norm, (b) the l_∞-vector norm.

Solution. We use (4*), Sec. 7.8, for the inverse and then (10) and (11) in Sec. 20.3. Thus

$$\mathbf{A} = \begin{bmatrix} 0.9999 & -1.0001 \\ 1.0000 & -1.0000 \end{bmatrix}, \qquad \mathbf{A}^{-1} = \begin{bmatrix} -5000.0 & 5000.5 \\ -5000.0 & 4999.5 \end{bmatrix}.$$

(a) The l_1-vector norm gives the column "sum" norm (10), Sec. 20.3; from Column 2 we thus obtain $\|\mathbf{A}\| = |-1.0001| + |-1.0000| = 2.0001$. Similarly, $\|\mathbf{A}^{-1}\| = 10{,}000$.

(b) The l_∞-vector norm gives the row "sum" norm (11), Sec. 20.3; thus $\|\mathbf{A}\| = 2$, $\|\mathbf{A}^{-1}\| = 10000.5$ from Row 1. We notice that $\|\mathbf{A}^{-1}\|$ is surprisingly large, which makes the product $\|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ large (20,001). We shall see below that this is typical of an ill-conditioned system.

Condition Number of a Matrix

We are now ready to introduce the key concept in our discussion of ill-conditioning, the **condition number** $\kappa(\mathbf{A})$ of a (nonsingular) square matrix $\mathbf{A}$, defined by

(13)
$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|.$$

The role of the condition number is seen from the following theorem.

THEOREM 1

Condition Number

A linear system of equations $\mathbf{Ax} = \mathbf{b}$ *and its matrix* $\mathbf{A}$ *whose condition number* (13) *is small are well-conditioned. A large condition number indicates ill-conditioning.*

PROOF $\mathbf{b} = \mathbf{Ax}$ and (11) give $\|\mathbf{b}\| \leqq \|\mathbf{A}\|\,\|\mathbf{x}\|$. Let $\mathbf{b} \neq \mathbf{0}$ and $\mathbf{x} \neq \mathbf{0}$. Then division by $\|\mathbf{b}\|\,\|\mathbf{x}\|$ gives

(14)
$$\frac{1}{\|\mathbf{x}\|} \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}.$$

Multiplying (2) $\mathbf{r} = \mathbf{A}(\mathbf{x} - \tilde{\mathbf{x}})$ by $\mathbf{A}^{-1}$ from the left and interchanging sides, we have $\mathbf{x} - \tilde{\mathbf{x}} = \mathbf{A}^{-1}\mathbf{r}$. Now (11) with $\mathbf{A}^{-1}$ and $\mathbf{r}$ instead of $\mathbf{A}$ and $\mathbf{x}$ yields

$$\|\mathbf{x} - \tilde{\mathbf{x}}\| = \|\mathbf{A}^{-1}\mathbf{r}\| \leqq \|\mathbf{A}^{-1}\|\,\|\mathbf{r}\|.$$

Division by $\|\mathbf{x}\|$ [note that $\|\mathbf{x}\| \neq 0$ by (3b)] and use of (14) finally gives

(15)
$$\frac{\|\mathbf{x} - \tilde{\mathbf{x}}\|}{\|\mathbf{x}\|} \leqq \frac{1}{\|\mathbf{x}\|}\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| \leqq \frac{\|\mathbf{A}\|}{\|\mathbf{b}\|}\|\mathbf{A}^{-1}\|\,\|\mathbf{r}\| = \kappa(\mathbf{A})\frac{\|\mathbf{r}\|}{\|\mathbf{b}\|}.$$

Hence if $\kappa(\mathbf{A})$ is small, a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ implies a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$, so that the system is well-conditioned. However, this does not hold if $\kappa(\mathbf{A})$ is large; then a small $\|\mathbf{r}\|/\|\mathbf{b}\|$ does not necessarily imply a small relative error $\|\mathbf{x} - \tilde{\mathbf{x}}\|/\|\mathbf{x}\|$. ■

EXAMPLE 5 **Condition Numbers. Gauss–Seidel Iteration**

$$\mathbf{A} = \begin{bmatrix} 5 & 1 & 1 \\ 1 & 4 & 2 \\ 1 & 2 & 4 \end{bmatrix} \quad \text{has the inverse} \quad \mathbf{A}^{-1} = \frac{1}{56}\begin{bmatrix} 12 & -2 & -2 \\ -2 & 19 & -9 \\ -2 & -9 & 19 \end{bmatrix}.$$

Since $\mathbf{A}$ is symmetric, (10) and (11) in Sec. 20.3 give the same condition number

$$\kappa(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\| = 7 \cdot \tfrac{1}{56} \cdot 30 = 3.75.$$

We see that a linear system $\mathbf{Ax} = \mathbf{b}$ with this $\mathbf{A}$ is well-conditioned.

For instance, if $\mathbf{b} = [14 \quad 0 \quad 28]^\mathsf{T}$, the Gauss algorithm gives the solution $\mathbf{x} = [2 \quad -5 \quad 9]^\mathsf{T}$, (confirm this). Since the main diagonal entries of $\mathbf{A}$ are relatively large, we can expect reasonably good convergence of the Gauss–Seidel iteration. Indeed, starting from, say, $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$, we obtain the first 8 steps (3D values)

x_1	x_2	x_3
1.000	1.000	1.000
2.400	−1.100	6.950
1.630	−3.882	8.534
1.870	−4.734	8.900
1.967	−4.942	8.979
1.993	−4.988	8.996
1.998	−4.997	8.999
2.000	−5.000	9.000
2.000	−5.000	9.000

■

EXAMPLE 6 Ill-Conditioned Linear System

Example 4 gives by (10) or (11), Sec. 20.3, for the matrix in Example 1 the very large condition number $\kappa(\mathbf{A}) = 2.0001 \cdot 10000 = 2 \cdot 10000.5 = 200001$. This confirms that the system is very ill-conditioned.

Similarly in Example 2, where by (4*), Sec. 7.8 and 6D-computation,

$$\mathbf{A}^{-1} = \frac{1}{0.0002}\begin{bmatrix} 1.0001 & -1.0000 \\ -1.0000 & 1.0001 \end{bmatrix} = \begin{bmatrix} 5000.5 & -5.000.0 \\ -5000.0 & 5000.5 \end{bmatrix}$$

so that (10), Sec. 20.3, gives a very large $\kappa(\mathbf{A})$, explaining the surprising result in Example 2,

$$\kappa(\mathbf{A}) = (1.0001 + 1.0000)(5000.5 + 5000.0) \approx 20{,}002.$$

■

In practice, $\mathbf{A}^{-1}$ will not be known, so that in computing the condition number $\kappa(\mathbf{A})$, one must estimate $\|\mathbf{A}^{-1}\|$. A method for this (proposed in 1979) is explained in Ref. [E9] listed in App. 1.

Inaccurate Matrix Entries. $\kappa(\mathbf{A})$ can be used for estimating the effect $\delta\mathbf{x}$ of an inaccuracy $\delta\mathbf{A}$ of $\mathbf{A}$ (errors of measurements of the a_{jk}, for instance). Instead of $\mathbf{Ax} = \mathbf{b}$ we then have

$$(\mathbf{A} + \delta\mathbf{A})(\mathbf{x} + \delta\mathbf{x}) = \mathbf{b}.$$

Multiplying out and subtracting $\mathbf{Ax} = \mathbf{b}$ on both sides, we obtain

$$\mathbf{A}\delta\mathbf{x} + \delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}) = \mathbf{0}.$$

Multiplication by $\mathbf{A}^{-1}$ from the left and taking the second term to the right gives

$$\delta\mathbf{x} = -\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x}).$$

Applying (11) with $\mathbf{A}^{-1}$ and vector $\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})$ instead of $\mathbf{A}$ and $\mathbf{x}$, we get

$$\|\delta\mathbf{x}\| = \|\mathbf{A}^{-1}\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}(\mathbf{x} + \delta\mathbf{x})\|.$$

Applying (11) on the right, with $\delta\mathbf{A}$ and $\mathbf{x} - \delta\mathbf{x}$ instead of $\mathbf{A}$ and $\mathbf{x}$, we obtain

$$\|\delta\mathbf{x}\| \leqq \|\mathbf{A}^{-1}\|\,\|\delta\mathbf{A}\|\,\|\mathbf{x} + \delta\mathbf{x}\|.$$

Now $\|\mathbf{A}^{-1}\| = \kappa(\mathbf{A})/\|\mathbf{A}\|$ by the definition of $\kappa(\mathbf{A})$, so that division by $\|\mathbf{x} + \delta\mathbf{x}\|$ shows that the relative inaccuracy of $\mathbf{x}$ is related to that of $\mathbf{A}$ via the condition number by the inequality

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \approx \frac{\|\delta\mathbf{x}\|}{\|\mathbf{x} + \delta\mathbf{x}\|} \leqq \|\mathbf{A}^{-1}\| \, \|\delta\mathbf{A}\| = \kappa(\mathbf{A}) \frac{\|\delta\mathbf{A}\|}{\|\mathbf{A}\|}. \tag{16}$$

Conclusion. If the system is well-conditioned, small inaccuracies $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ can have only a small effect on the solution. However, in the case of ill-conditioning, if $\|\delta\mathbf{A}\|/\|\mathbf{A}\|$ is small, $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ ***may*** be large.

Inaccurate Right Side. You may show that, similarly, when $\mathbf{A}$ is accurate, an inaccuracy $\delta\mathbf{b}$ of $\mathbf{b}$ causes an inaccuracy $\delta\mathbf{x}$ satisfying

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq \kappa(\mathbf{A}) \frac{\|\delta\mathbf{b}\|}{\|\mathbf{b}\|}. \tag{17}$$

Hence $\|\delta\mathbf{x}\|/\|\mathbf{x}\|$ must remain relatively small whenever $\kappa(\mathbf{A})$ is small.

EXAMPLE 7 Inaccuracies. Bounds (16) and (17)

If each of the nine entries of $\mathbf{A}$ in Example 5 is measured with an inaccuracy of 0.1, then $\|\delta\mathbf{A}\| = 9 \cdot 0.1$ and (16) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{3 \cdot 0.1}{7} = 0.321 \qquad \text{thus} \qquad \|\delta\mathbf{x}\| \leqq 0.321\,\|\mathbf{x}\| = 0.321 \cdot 16 = 5.14.$$

By experimentation you will find that the actual inaccuracy $\|\delta\mathbf{x}\|$ is only about 30% of the bound 5.14. This is typical.

Similarly, if $\delta\mathbf{b} = [0.1 \quad 0.1 \quad 0.1]^{\mathsf{T}}$, then $\|\delta\mathbf{b}\| = 0.3$ and $\|\mathbf{b}\| = 42$ in Example 5, so that (17) gives

$$\frac{\|\delta\mathbf{x}\|}{\|\mathbf{x}\|} \leqq 7.5 \cdot \frac{0.3}{42} = 0.0536, \qquad \text{hence} \qquad \|\delta\mathbf{x}\| \leqq 0.0536 \cdot 16 = 0.857$$

but this bound is again much greater than the actual inaccuracy, which is about 0.15. ■

Further Comments on Condition Numbers. The following additional explanations may be helpful.

1. There is no sharp dividing line between "well-conditioned" and "ill-conditioned," but generally the situation will get worse as we go from systems with small $\kappa(\mathbf{A})$ to systems with larger $\kappa(\mathbf{A})$. Now always $\kappa(\mathbf{A}) \geqq 1$, so that values of 10 or 20 or so give no reason for concern, whereas $\kappa(\mathbf{A}) = 100$, say, calls for caution, and systems such as those in Examples 1 and 2 are extremely ill-conditioned.

2. If $\kappa(\mathbf{A})$ is large (or small) in one norm, it will be large (or small, respectively) in any other norm. See Example 5.

3. The literature on ill-conditioning is extensive. For an introduction to it, see [E9].

This is the end of our discussion of numerics for solving linear systems. In the next section we consider curve fitting, an important area in which solutions are obtained from linear systems.

PROBLEM SET 20.4

1–6 VECTOR NORMS

Compute the norms (5), (6), (7). Compute a corresponding **unit vector** (vector of norm 1) with respect to the l_∞-norm.

1. $[1 \quad -3 \quad 8 \quad 0 \quad -6 \quad 0]$
2. $[4 \quad -1 \quad 8]$
3. $[0.2 \quad 0.6 \quad -2.1 \quad 3.0]$
4. $[k^2, \quad 4k, \quad k^3], \quad k > 4$
5. $[1 \quad 1 \quad 1 \quad 1 \quad 1]$
6. $[0 \quad 0 \quad 0 \quad 1 \quad 0]$
7. For what $\mathbf{x} = [a \quad b \quad c]$ will $\|\mathbf{x}\|_1 = \|\mathbf{x}\|_2$?
8. Show that $\|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1$.

9–16 MATRIX NORMS, CONDITION NUMBERS

Compute the matrix norm and the condition number corresponding to the l_1-vector norm.

9. $\begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix}$ **10.** $\begin{bmatrix} 2.1 & 4.5 \\ 0.5 & 1.8 \end{bmatrix}$

11. $\begin{bmatrix} \sqrt{5} & 5 \\ 0 & -\sqrt{5} \end{bmatrix}$ **12.** $\begin{bmatrix} 7 & 6 \\ 6 & 5 \end{bmatrix}$

13. $\begin{bmatrix} -2 & 4 & -1 \\ -2 & 3 & 0 \\ 7 & -12 & 2 \end{bmatrix}$ **14.** $\begin{bmatrix} 1 & 0.01 & 0 \\ 0.01 & 1 & 0.01 \\ 0 & 0.01 & 1 \end{bmatrix}$

15. $\begin{bmatrix} -20 & 0 & 0 \\ 0 & 0.05 & 0 \\ 0 & 0 & 20 \end{bmatrix}$

16. $\begin{bmatrix} 21 & 10.5 & 7 & 5.25 \\ 10.5 & 7 & 5.25 & 4.2 \\ 7 & 5.25 & 4.2 & 3.5 \\ 5.25 & 4.2 & 3.5 & 3 \end{bmatrix}$

17. Verify (11) for $\mathbf{x} = [3 \quad 15 \quad -4]^\mathsf{T}$ taken with the l_∞-norm and the matrix in Prob. 13.

18. Verify (12) for the matrices in Probs. 9 and 10.

19–20 ILL-CONDITIONED SYSTEMS

Solve $\mathbf{Ax} = \mathbf{b}_1$, $\mathbf{Ax} = \mathbf{b}_2$. Compare the solutions and comment. Compute the condition number of $\mathbf{A}$.

19. $\mathbf{A} = \begin{bmatrix} 4.50 & 3.55 \\ 3.55 & 2.80 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 5.2 \\ 4.1 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 5.2 \\ 4.0 \end{bmatrix}$

20. $\mathbf{A} = \begin{bmatrix} 3.0 & 1.7 \\ 1.7 & 1.0 \end{bmatrix}, \quad \mathbf{b}_1 = \begin{bmatrix} 4.7 \\ 2.7 \end{bmatrix}, \quad \mathbf{b}_2 = \begin{bmatrix} 4.7 \\ 2.71 \end{bmatrix}$

21. Residual. For $\mathbf{Ax} = \mathbf{b}_1$ in Prob. 19 guess what the residual of $\tilde{\mathbf{x}} = [-10.0 \quad 14.1]^\mathsf{T}$, very poorly approximating $[-2 \quad 4]^\mathsf{T}$, might be. Then calculate and comment.

22. Show that $\kappa(\mathbf{A}) \geqq 1$ for the matrix norms (10), (11), Sec. 20.3, and $\kappa(\mathbf{A}) \geqq \sqrt{n}$ for the Frobenius norm (9), Sec. 20.3.

23. CAS EXPERIMENT. Hilbert Matrices. The 3×3 Hilbert matrix is

$$\mathbf{H}_3 = \begin{bmatrix} 1 & \frac{1}{2} & \frac{1}{3} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \end{bmatrix}.$$

The $n \times n$ Hilbert matrix is $\mathbf{H}_n = [h_{jk}]$, where $h_{jk} = 1/(j + k - 1)$. (Similar matrices occur in curve fitting by least squares.) Compute the condition number $\kappa(\mathbf{H}_n)$ for the matrix norm corresponding to the l_∞- (or l_1-) vector norm, for $n = 2, 3, \cdots, 6$ (or further if you wish). Try to find a formula that gives reasonable approximate values of these rapidly growing numbers.

Solve a few linear systems of your choice, involving an $\mathbf{H}_n$.

24. TEAM PROJECT. Norms. (a) Vector norms in our text are **equivalent**, that is, they are related by double inequalities; for instance,

$$\text{(18)} \quad \begin{aligned} &\text{(a)} \quad \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1 \leqq n\|\mathbf{x}\|_\infty \\ &\text{(b)} \quad \frac{1}{n}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_\infty \leqq \|\mathbf{x}\|_1. \end{aligned}$$

Hence if for some $\mathbf{x}$, one norm is large (or small), the other norm must also be large (or small). Thus in many investigations the particular choice of a norm is not essential. Prove (18).

(b) The Cauchy–Schwarz inequality is

$$|\mathbf{x}^\mathsf{T}\mathbf{y}| \leqq \|\mathbf{x}\|_2 \|\mathbf{y}\|_2.$$

It is very important. (Proof in Ref. [GenRef7] listed in App. 1.) Use it to prove

(19a) $$\|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1 \leqq \sqrt{n}\,\|\mathbf{x}\|_2$$

(19b) $$\frac{1}{\sqrt{n}}\|\mathbf{x}\|_1 \leqq \|\mathbf{x}\|_2 \leqq \|\mathbf{x}\|_1.$$

(c) Formula (10) is often more practical than (9). Derive (10) from (9).

(d) Matrix norms. Illustrate (11) with examples. Give examples of (12) with equality as well as with strict inequality. Prove that the matrix norms (10), (11) in Sec. 20.3 satisfy the *axioms of a norm*

$$\|\mathbf{A}\| \geqq \mathbf{0}.$$
$$\|\mathbf{A}\| = \mathbf{0} \text{ if and only if } \mathbf{A} = \mathbf{0},$$
$$\|k\mathbf{A}\| = |k|\,\|\mathbf{A}\|,$$
$$\|\mathbf{A} + \mathbf{B}\| \leqq \|\mathbf{A}\| + \|\mathbf{B}\|.$$

25. WRITING PROJECT. Norms and Their Use in This Section. Make a list of the most important of the many ideas covered in this section and write a two-page report on them.

20.5 Least Squares Method

Having discussed numerics for linear systems, we now turn to an important application, curve fitting, in which the solutions are obtained from linear systems.

In **curve fitting** we are given n points (pairs of numbers) $(x_1, y_1), \cdots, (x_n, y_n)$ and we want to determine a function $f(x)$ such that

$$f(x_1) \approx y_1, \cdots, f(x_n) \approx y_n,$$

approximately. The type of function (for example, polynomials, exponential functions, sine and cosine functions) may be suggested by the nature of the problem (the underlying physical law, for instance), and in many cases a polynomial of a certain degree will be appropriate.

Let us begin with a motivation.

If we require strict equality $f(x_1) = y_1, \cdots, f(x_n) = y_n$ and use polynomials of sufficiently high degree, we may apply one of the methods discussed in Sec. 19.3 in connection with interpolation. However, in certain situations this would not be the appropriate solution of the actual problem. For instance, to the four points

(1) $$(-1.3,\ 0.103), \qquad (-0.1,\ 1.099), \qquad (0.2,\ 0.808), \qquad (1.3,\ 1.897)$$

there corresponds the interpolation polynomial $f(x) = x^3 - x + 1$ (Fig. 446), but if we graph the points, we see that they lie nearly on a straight line. Hence if these values are obtained in an experiment and thus involve an experimental error, and if the nature of the experiment suggests a linear relation, we better fit a straight line through the points (Fig. 446). Such a line may be useful for predicting values to be expected for other values of x. A widely used principle for fitting straight lines is the **method**

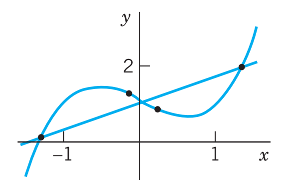


Fig. 446. Approximate fitting of a straight line

of least squares by Gauss and Legendre. In the present situation it may be formulated as follows.

Method of Least Squares. *The straight line*

$$y = a + bx \tag{2}$$

should be fitted through the given points $(x_1, y_1), \cdots, (x_n, y_n)$ *so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction* (*the* y-*direction*).

The point on the line with abscissa x_j has the ordinate $a + bx_j$. Hence its distance from (x_j, y_j) is $|y_j - a - bx_j|$ (Fig. 447) and that sum of squares is

$$q = \sum_{j=1}^{n} (y_j - a - bx_j)^2.$$

q depends on a and b. A necessary condition for q to be minimum is

$$\begin{aligned} \frac{\partial q}{\partial a} &= -2 \sum (y_j - a - bx_j) = 0 \\ \frac{\partial q}{\partial b} &= -2 \sum x_j (y_j - a - bx_j) = 0 \end{aligned} \tag{3}$$

(where we sum over j from 1 to n). Dividing by 2, writing each sum as three sums, and taking one of them to the right, we obtain the result

$$\begin{aligned} an \quad + b \sum x_j &= \sum y_j \\ a \sum x_j + b \sum x_j^2 &= \sum x_j y_j. \end{aligned} \tag{4}$$

These equations are called the **normal equations** of our problem.

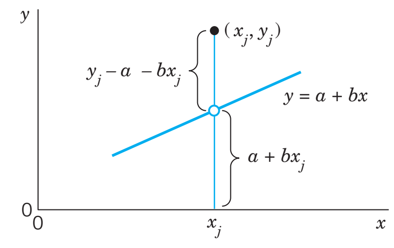


Fig. 447. Vetrical distance of a point (x_j, y_j) from a straight line $y = a + bx$

EXAMPLE 1 Straight Line

Using the method of least squares, fit a straight line to the four points given in formula (1).

Solution. We obtain

$$n = 4, \qquad \sum x_j = 0.1, \qquad \sum x_j^2 = 3.43, \qquad \sum y_j = 3.907, \qquad \sum x_j y_j = 2.3839.$$

Hence the normal equations are

$$4a + 0.10b = 3.9070$$
$$0.1a + 3.43b = 2.3839.$$

The solution (rounded to 4D) is $a = 0.9601$, $b = 0.6670$, and we obtain the straight line (Fig. 446)

$$y = 0.9601 + 0.6670x.$$

Curve Fitting by Polynomials of Degree m

Our method of curve fitting can be generalized from a polynomial $y = a + bx$ to a polynomial of degree m

$$p(x) = b_0 + b_1 x + \cdots + b_m x^m \tag{5}$$

where $m \leqq n - 1$. Then q takes the form

$$q = \sum_{j=1}^{n} (y_j - p(x_j))^2$$

and depends on $m + 1$ parameters $b_0, \cdots, b_m$. Instead of (3) we then have $m + 1$ conditions

$$\frac{\partial q}{\partial b_0} = 0, \qquad \cdots, \qquad \frac{\partial q}{\partial b_m} = 0 \tag{6}$$

which give a system of $m + 1$ normal equations.

In the case of a quadratic polynomial

$$p(x) = b_0 + b_1 x + b_2 x^2 \tag{7}$$

the normal equations are (summation from 1 to n)

$$\begin{aligned} b_0 n \quad &+ b_1 \sum x_j + b_2 \sum x_j^2 = \sum y_j \\ b_0 \sum x_j &+ b_1 \sum x_j^2 + b_2 \sum x_j^3 = \sum x_j y_j \\ b_0 \sum x_j^2 &+ b_1 \sum x_j^3 + b_2 \sum x_j^4 = \sum x_j^2 y_j. \end{aligned} \tag{8}$$

The derivation of (8) is left to the reader.

EXAMPLE 2 Quadratic Parabola by Least Squares

Fit a parabola through the data (0, 5), (2, 4), (4, 1), (6, 6), (8, 7).

Solution. For the normal equations we need $n = 5$, $\Sigma x_j = 20$, $\Sigma x_j^2 = 120$, $\Sigma x_j^3 = 800$, $\Sigma x_j^4 = 5664$, $\Sigma y_j = 23$, $\Sigma x_j y_j = 104$, $\Sigma x_j^2 y_j = 696$. Hence these equations are

$$\begin{aligned} 5b_0 + 20b_1 + 120b_2 &= 23 \\ 20b_0 + 120b_1 + 800b_2 &= 104 \\ 120b_0 + 800b_1 + 5664b_2 &= 696. \end{aligned}$$

Solving them we obtain the quadratic least squares parabola (Fig. 448)

$$y = 5.11429 - 1.41429x + 0.21429x^2.$$

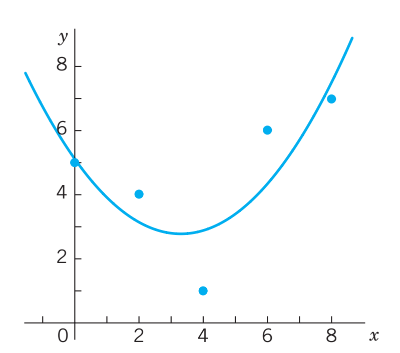


Fig. 448. Least squares parabola in Example 2

For a general polynomial (5) the normal equations form a linear system of equations in the unknowns $b_0, \cdots, b_m$. When its matrix $\mathbf{M}$ is nonsingular, we can solve the system by Cholesky's method (Sec. 20.2) because then $\mathbf{M}$ is positive definite (and symmetric). When the equations are nearly linearly dependent, the normal equations may become ill-conditioned and should be replaced by other methods; see [E5], Sec. 5.7, listed in App. 1.

The least squares method also plays a role in statistics (see Sec. 25.9).

PROBLEM SET 20.5

1–6 FITTING A STRAIGHT LINE

Fit a straight line to the given points (x, y) by least squares. Show the details. Check your result by sketching the points and the line. Judge the goodness of fit.

1. $(0, 2)$, $(2, 0)$, $(3, -2)$, $(5, -3)$
2. How does the line in Prob. 1 change if you add a point far above it, say, $(1, 3)$? Guess first.
3. $(0, 1.8)$, $(1, 1.6)$, $(2, 1.1)$, $(3, 1.5)$, $(4, 2.3)$
4. **Hooke's law $F = ks$.** Estimate the spring modulus k from the force F [lb] and the elongation s [cm], where $(F, s) = (1, 0.3)$, $(2, 0.7)$, $(4, 1.3)$, $(6, 1.9)$, $(10, 3.2)$, $(20, 6.3)$.
5. **Average speed.** Estimate the average speed v_{av} of a car traveling according to $s = v \cdot t$ [km] (s = distance traveled, t [hr] = time) from $(t, s) = (9, 140)$, $(10, 220)$, $(11, 310)$, $(12, 410)$.
6. **Ohm's law $U = Ri$.** Estimate R from $(i, U) = (2, 104)$, $(4, 206)$, $(6, 314)$, $(10, 530)$.
7. Derive the normal equations (8).

8–11 FITTING A QUADRATIC PARABOLA

Fit a parabola (7) to the points (x, y). Check by sketching.

8. $(-1, 5)$, $(1, 3)$, $(2, 4)$, $(3, 8)$
9. $(2, -3)$, $(3, 0)$, $(5, 1)$, $(6, 0)$ $(7, -2)$
10. t [hr] = Worker's time on duty, y [sec] = His/her reaction time, $(t, y) = (1, 2.0)$, $(2, 1.78)$, $(3, 1.90)$, $(4, 2.35)$, $(5, 2.70)$
11. The data in Prob. 3. Plot the points, the line, and the parabola jointly. Compare and comment.
12. **Cubic parabola.** Derive the formula for the normal equations of a cubic least squares parabola.
13. Fit curves (2) and (7) and a cubic parabola by least squares to $(x, y) = (-2, -30)$, $(-1, -4)$, $(0, 4)$, $(1, 4)$, $(2, 22)$, $(3, 68)$. Graph these curves and the points on common axes. Comment on the goodness of fit.
14. **TEAM PROJECT.** The **least squares approximation of a function** $f(x)$ on an interval $a \leqq x \leqq b$ by a function

$$F_m(x) = a_0 y_0(x) + a_1 y_1(x) + \cdots + a_m y_m(x)$$

where $y_0(x), \cdots, y_m(x)$ are given functions, requires the determination of the coefficients $a_0, \cdots, a_m$ such that

$$\int_a^b [f(x) - F_m(x)]^2 \, dx \tag{9}$$

becomes minimum. This integral is denoted by $\|f - F_m\|^2$, and $\|f - F_m\|$ is called the **L_2-norm** of $f - F_m$ (L suggesting Lebesgue[5]). A necessary condition for that minimum is given by $\partial\|f - F_m\|^2/\partial a_j = 0$, $j = 0, \cdots, m$ [the analog of (6)]. **(a)** Show that this leads to $m + 1$ normal equations ($j = 0, \cdots, m$)

$$\sum_{k=0}^{m} h_{jk} a_k = b_j \qquad \text{where}$$

$$h_{jk} = \int_a^b y_j(x) y_k(x) \, dx, \tag{10}$$

$$b_j = \int_a^b f(x) y_j(x) \, dx.$$

(b) Polynomial. What form does (10) take if $F_m(x) = a_0 + a_1 x + \cdots + a_m x^m$? What is the coefficient matrix of (10) in this case when the interval is $0 \leqq x \leqq 1$?

(c) Orthogonal functions. What are the solutions of (10) if $y_0(x), \cdots, y_m(x)$ are orthogonal on the interval $a \leqq x \leqq b$? (For the definition, see Sec. 11.5. See also Sec. 11.6.)

15. CAS EXPERIMENT. Least Squares versus Interpolation. For the given data and for data of your choice find the interpolation polynomial and the least squares approximations (linear, quadratic, etc.). Compare and comment.

(a) $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$

(b) $(-4, 0)$, $(-3, 0)$, $(-2, 0)$, $(-1, 0)$, $(0, 1)$, $(1, 0)$, $(2, 0)$, $(3, 0)$, $(4, 0)$

(c) Choose five points on a straight line, e.g., $(0, 0)$, $(1, 1), \cdots, (4, 4)$. Move one point 1 unit upward and find the quadratic least squares polynomial. Do this for each point. Graph the five polynomials on common axes. Which of the five motions has the greatest effect?

20.6 Matrix Eigenvalue Problems: Introduction

We now come to the second part of our chapter on numeric linear algebra. In the *first part of this chapter* we discussed methods of solving systems of linear equations, which included Gauss elimination with backward substitution. This method is known as a direct method since it gives solutions after a prescribed amount of computation. The Gauss method was modified by Doolittle's method, Crout's method, and Cholesky's method, each requiring fewer arithmetic operations than Gauss. Finally we presented indirect methods of solving systems of linear equations, that is, the Gauss–Seidel method and the Jacobi iteration. The indirect methods require an undetermined number of iterations. That number depends on how far we start from the true solution and what degree of accuracy we require. Moreover, depending on the problem, convergence may be fast or slow or our computation cycle might not even converge. This led to the concepts of ill-conditioned problems and condition numbers that help us gain some control over difficulties inherent in numerics.

The second part of this chapter deals with some of the most important ideas and numeric methods for matrix eigenvalue problems. This very extensive part of numeric linear algebra is of great practical importance, with much research going on, and hundreds, if not thousands, of papers published in various mathematical journals (see the references in [E8], [E9], [E11], [E29]). We begin with the concepts and general results we shall need in explaining and applying numeric methods for eigenvalue problems. (For typical models of eigenvalue problems see Chap. 8.)

[5]HENRI LEBESGUE (1875–1941), great French mathematician, creator of a modern theory of measure and integration in his famous doctoral thesis of 1902.

An **eigenvalue** or **characteristic value** (or *latent root*) of a given $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is a real or complex number λ such that the vector equation

$$\text{(1)} \qquad \mathbf{Ax} = \lambda \mathbf{x}$$

has a nontrivial solution, that is, a solution $\mathbf{x} \neq \mathbf{0}$, which is then called an **eigenvector** or **characteristic vector** of $\mathbf{A}$ corresponding to that eigenvalue λ. The set of all eigenvalues of $\mathbf{A}$ is called the **spectrum** of **A.** Equation (1) can be written

$$\text{(2)} \qquad (\mathbf{A} - \lambda \mathbf{I})\mathbf{x} = \mathbf{0}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix. This homogeneous system has a nontrivial solution if and only if the **characteristic determinant** det $(\mathbf{A} - \lambda\mathbf{I})$ is 0 (see Theorem 2 in Sec. 7.5). This gives (see Sec. 8.1)

THEOREM 1

Eigenvalues

The eigenvalues of **A** *are the solutions* λ *of the* **characteristic equation**

$$\text{(3)} \qquad \det(\mathbf{A} - \lambda\mathbf{I}) = \begin{vmatrix} a_{11} - \lambda & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} - \lambda & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} - \lambda \end{vmatrix} = 0.$$

Developing the characteristic determinant, we obtain the **characteristic polynomial** of **A**, which is of degree n in λ. Hence **A** has at least one and at most n numerically different eigenvalues. If **A** is real, so are the coefficients of the characteristic polynomial. By familiar algebra it follows that then the roots (the eigenvalues of **A**) are ***real or complex conjugates*** in pairs.

To give you some orientation of the underlying approaches of numerics for eigenvalue problems, note the following. For large or very large matrices it may be very difficult to determine the eigenvalues, since, in general, it is difficult to find the roots of characteristic polynomials of higher degrees. We will discuss different numeric methods for finding eigenvalues that achieve different results. Some methods, such as in Sec. 20.7, will give us only regions in which complex eigenvalues lie (Geschgorin's method) or the intervals in which the largest and smallest real eigenvalue lie (Collatz method). Other methods compute all eigenvalues, such as the Householder tridiagonalization method and the QR-method in Sec. 20.9.

To continue our discussion, we shall usually denote the eigenvalues of **A** by

$$\lambda_1, \lambda_2, \cdots, \lambda_n$$

with the understanding that some (or all) of them may be equal.

The sum of these n eigenvalues equals the sum of the entries on the main diagonal of **A**, called the trace of **A**; thus

$$\text{(4)} \qquad \text{trace}\,\mathbf{A} = \sum_{j=1}^{n} a_{jj} = \sum_{k=1}^{n} \lambda_k.$$

Also, the product of the eigenvalues equals the determinant of $\mathbf{A}$,

$$\det \mathbf{A} = \lambda_1 \lambda_2 \cdots \lambda_n. \tag{5}$$

Both formulas follow from the product representation of the characteristic polynomial, which we denote by $f(\lambda)$,

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)(\lambda - \lambda_2) \cdots (\lambda - \lambda_n).$$

If we take equal factors together and denote the *numerically distinct* eigenvalues of $\mathbf{A}$ by $\lambda_1, \cdots, \lambda_r$ $(r \leqq n)$, then the product becomes

$$f(\lambda) = (-1)^n (\lambda - \lambda_1)^{m_1} (\lambda - \lambda_2)^{m_2} \cdots (\lambda - \lambda_r)^{m_r}. \tag{6}$$

The exponent m_j is called the **algebraic multiplicity** of λ_j. The maximum number of linearly independent eigenvectors corresponding to λ_j is called the **geometric multiplicity** of λ_j. It is equal to or smaller than m_j.

A subspace S of R^n or C^n (if $\mathbf{A}$ is complex) is called an **invariant subspace** of $\mathbf{A}$ if for every $\mathbf{v}$ in S the vector $\mathbf{Av}$ is also in S. **Eigenspaces** of $\mathbf{A}$ (spaces of eigenvectors; Sec. 8.1) are important invariant subspaces of $\mathbf{A}$.

An $n \times n$ matrix $\mathbf{B}$ is called **similar** to $\mathbf{A}$ if there is a nonsingular $n \times n$ matrix $\mathbf{T}$ such that

$$\mathbf{B} = \mathbf{T}^{-1} \mathbf{A} \mathbf{T}. \tag{7}$$

Similarity is important for the following reason.

THEOREM 2

Similar Matrices

Similar matrices have the same eigenvalues. If $\mathbf{x}$ *is an eigenvector of* $\mathbf{A}$, *then* $\mathbf{y} = \mathbf{T}^{-1}\mathbf{x}$ *is an eigenvector of* $\mathbf{B}$ *in* (7) *corresponding to the same eigenvalue.* (Proof in Sec. 8.4.)

Another theorem that has various applications in numerics is as follows.

THEOREM 3

Spectral Shift

If $\mathbf{A}$ *has the eigenvalues* $\lambda_1, \cdots, \lambda_n$, *then* $\mathbf{A} - k\mathbf{I}$ *with arbitrary k has the eigenvalues* $\lambda_1 - k, \cdots, \lambda_n - k$.

This theorem is a special case of the following **spectral mapping theorem**.

THEOREM 4

Polynomial Matrices

If λ *is an eigenvalue of* $\mathbf{A}$, *then*

$$q(\lambda) = \alpha_s \lambda^s + \alpha_{s-1} \lambda^{s-1} + \cdots + \alpha_1 \lambda + \alpha_0$$

is an eigenvalue of the **polynomial matrix**

$$q(\mathbf{A}) = \alpha_s \mathbf{A}^s + \alpha_{s-1} \mathbf{A}^{s-1} + \cdots + \alpha_1 \mathbf{A} + a_0 \mathbf{I}.$$

PROOF $\mathbf{Ax} = \lambda\mathbf{x}$ implies $\mathbf{A}^2\mathbf{x} = \mathbf{A}\lambda\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, $\mathbf{A}^3\mathbf{x} = \lambda^3\mathbf{x}$, etc. Thus

$$\begin{aligned} q(\mathbf{A})\mathbf{x} &= (\alpha_s\mathbf{A}^s + \alpha_{s-1}\mathbf{A}^{s-1} + \cdots)\,\mathbf{x} \\ &= \alpha_s\mathbf{A}^s\mathbf{x} + \alpha_{s-1}A^{s-1}\mathbf{x} + \cdots \\ &= \alpha_s\lambda^s\mathbf{x} + \alpha_{s-1}\lambda^{s-1}\mathbf{x} + \cdots = q(\lambda)\,\mathbf{x}. \end{aligned}$$ ■

The eigenvalues of important special matrices can be characterized as follows.

THEOREM 5

Special Matrices

The eigenvalues of Hermitian matrices (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}$), *hence of real symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}$), *are real. The eigenvalues of skew-Hermitian matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = -\mathbf{A}$), *hence of real skew-symmetric matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = -\mathbf{A}$), *are pure imaginary or* 0. *The eigenvalues of unitary matrices* (i.e., $\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{A}^{-1}$), *hence of orthogonal matrices* (i.e., $\mathbf{A}^{\mathsf{T}} = \mathbf{A}^{-1}$), *have absolute value* 1. (Proofs in Secs. 8.3 and 8.5.)

The **choice of a numeric method** for matrix eigenvalue problems depends essentially on two circumstances, on the kind of matrix (real symmetric, real general, complex, sparse, or full) and on the kind of information to be obtained, that is, whether one wants to know all eigenvalues or merely specific ones, for instance, the largest eigenvalue, whether eigenvalues *and* eigenvectors are wanted, and so on. It is clear that we cannot enter into a systematic discussion of all these and further possibilities that arise in practice, but we shall concentrate on some basic aspects and methods that will give us a general understanding of this fascinating field.

20.7 Inclusion of Matrix Eigenvalues

The whole of numerics for matrix eigenvalues is motivated by the fact that, except for a few trivial cases, we cannot determine eigenvalues *exactly* by a finite process because these values are the roots of a polynomial of nth degree. Hence we must mainly use iteration.

In this section we state a few general theorems that give approximations and error bounds for eigenvalues. Our matrices will continue to be real (except in formula (5) below), but since (nonsymmetric) matrices may have complex eigenvalues, complex numbers will play a (very modest) role in this section.

The important theorem by Gerschgorin gives a region consisting of closed circular disks in the complex plane and including all the eigenvalues of a given matrix. Indeed, for each $j = 1, \cdots, n$ the inequality (1) in the theorem determines a closed circular disk in the complex λ-plane with center a_{jj} and radius given by the right side of (1); and Theorem 1 states that each of the eigenvalues of $\mathbf{A}$ lies in one of these n disks.

THEOREM 1

Gerschgorin's Theorem[6]

Let λ *be an eigenvalue of an arbitrary* $n \times n$ *matrix* $\mathbf{A} = [a_{jk}]$. *Then for some integer* j $(1 \leqq j \leqq n)$ *we have*

$$(1) \qquad |a_{jj} - \lambda| \leqq |a_{j1}| + |a_{j2}| + \cdots + |a_{j,j-1}| + |a_{j,j+1}| + \cdots + |a_{jn}|.$$

---

[6]SEMYON ARANOVICH GERSCHGORIN (1901–1933), Russian mathematician.

PROOF Let $\mathbf{x}$ be an eigenvector corresponding to an eigenvalue λ of $\mathbf{A}$. Then

$$(2) \qquad \mathbf{Ax} = \lambda\mathbf{x} \qquad \text{or} \qquad (\mathbf{A} - \lambda\mathbf{I})\mathbf{x} = \mathbf{0}.$$

Let x_j be a component of $\mathbf{x}$ that is largest in absolute value. Then we have $|x_m/x_j| \leqq 1$ for $m = 1, \cdots, n$. The vector equation (2) is equivalent to a system of n equations for the n components of the vectors on both sides. The jth of these n equations with j as just indicated is

$$a_{j1}x_1 + \cdots + a_{j,j-1}x_{j-1} + (a_{jj} - \lambda)x_j + a_{j,j+1}x_{j+1} + \cdots + a_{jn}x_n = 0.$$

Division by x_j (which cannot be zero; why?) and reshuffling terms gives

$$a_{jj} - \lambda = -a_{j1}\frac{x_1}{x_j} - \cdots - a_{j,j-1}\frac{x_{j-1}}{x_j} - a_{j,j+1}\frac{x_{j+1}}{x_j} - \cdots - a_{jn}\frac{x_n}{x_j}.$$

By taking absolute values on both sides of this equation, applying the triangle inequality $|a + b| \leqq |a| + |b|$ (where a and b are any complex numbers), and observing that because of the choice of j (which is crucial!), $|x_1/x_j| \leqq 1, \cdots, |x_n/x_j| \leqq 1$, we obtain (1), and the theorem is proved. ■

EXAMPLE 1 **Gerschgorin's Theorem**

For the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & 5 & 1 \\ \frac{1}{2} & 1 & 1 \end{bmatrix}$$

we get the Gerschgorin disks (Fig. 449)

$$D_1\text{: Center } 0, \text{ radius } 1, \qquad D_2\text{: Center } 5, \text{ radius } 1.5, \qquad D_3\text{: Center } 1, \text{ radius } 1.5.$$

The centers are the main diagonal entries of $\mathbf{A}$. These would be the eigenvalues of $\mathbf{A}$ if $\mathbf{A}$ were diagonal. We can take these values as crude approximations of the unknown eigenvalues (3D-values) $\lambda_1 = -0.209$, $\lambda_2 = 5.305$, $\lambda_3 = 0.904$ (verify this); then the radii of the disks are corresponding error bounds.

Since $\mathbf{A}$ is symmetric, it follows from Theorem 5, Sec. 20.6, that the spectrum of $\mathbf{A}$ must actually lie in the intervals $[-1, 2.5]$ and $[3.5, 6.5]$.

It is interesting that here the Gerschgorin disks form two disjoint sets, namely, $D_1 \cup D_3$, which contains two eigenvalues, and D_2, which contains one eigenvalue. This is typical, as the following theorem shows. ■

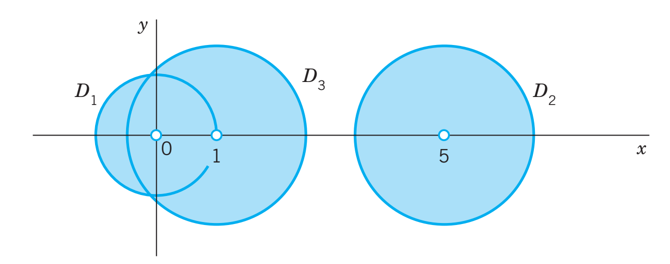


Fig. 449. Gerschgorin disks in Example 1

THEOREM 2

Extension of Gerschgorin's Theorem

If p Gerschgorin disks form a set S that is disjoint from the $n - p$ other disks of a given matrix $\mathbf{A}$, then S contains precisely p eigenvalues of $\mathbf{A}$ (each counted with its algebraic multiplicity, as defined in Sec. 20.6).

Idea of Proof. Set $\mathbf{A} = \mathbf{B} + \mathbf{C}$, where $\mathbf{B}$ is the diagonal matrix with entries a_{jj}, and apply Theorem 1 to $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$ with real t growing from 0 to 1. ■

EXAMPLE 2

Another Application of Gerschgorin's Theorem. Similarity

Suppose that we have diagonalized a matrix by some numeric method that left us with some off-diagonal entries of size 10^{-5}, say,

$$\mathbf{A} = \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix}.$$

What can we conclude about deviations of the eigenvalues from the main diagonal entries?

Solution. By Theorem 2, one eigenvalue must lie in the disk of radius $2 \cdot 10^{-5}$ centered at 4 and two eigenvalues (or an eigenvalue of algebraic multiplicity 2) in the disk of radius $2 \cdot 10^{-5}$ centered at 2. Actually, since the matrix is symmetric, these eigenvalues must lie in the intersections of these disks and the real axis, by Theorem 5 in Sec. 20.6.

We show how an isolated disk can always be reduced in size by a similarity transformation. The matrix

$$\mathbf{B} = \mathbf{T}^{-1}\mathbf{A}\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{-5} \end{bmatrix} \begin{bmatrix} 2 & 10^{-5} & 10^{-5} \\ 10^{-5} & 2 & 10^{-5} \\ 10^{-5} & 10^{-5} & 4 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 10^{5} \end{bmatrix}$$
$$= \begin{bmatrix} 2 & 10^{-5} & 1 \\ 10^{-5} & 2 & 1 \\ 10^{-10} & 10^{-10} & 4 \end{bmatrix}$$

is similar to $\mathbf{A}$. Hence by Theorem 2, Sec. 20.6, it has the same eigenvalues as $\mathbf{A}$. From Row 3 we get the smaller disk of radius $2 \cdot 10^{-10}$. Note that the other disks got bigger, approximately by a factor of 10^5. And in choosing $\mathbf{T}$ we have to watch that the new disks do not overlap with the disk whose size we want to decrease.

For further interesting facts, see the book [E28]. ■

By definition, a **diagonally dominant** matrix $\mathbf{A} = [a_{jk}]$ is an $n \times n$ matrix such that

$$|a_{jj}| \geqq \sum_{k \neq j} |a_{jk}| \qquad j = 1, \cdots, n \tag{3}$$

where we sum over all off-diagonal entries in Row j. The matrix is said to be **strictly diagonally dominant** if $>$ in (3) for all j. Use Theorem 1 to prove the following basic property.

THEOREM 3

Strict Diagonal Dominance

Strictly diagonally dominant matrices are nonsingular.

Further Inclusion Theorems

An **inclusion theorem** is a theorem that specifies a set which contains at least one eigenvalue of a given matrix. Thus, Theorems 1 and 2 are inclusion theorems; they even include the whole spectrum. We now discuss some famous theorems that yield further inclusions of eigenvalues. We state the first two of them without proofs (which would exceed the level of this book).

THEOREM 4

Schur's Theorem[7]

Let $\mathbf{A} = [a_{jk}]$ *be a* $n \times n$ *matrix. Then for each of its eigenvalues* $\lambda_1, \cdots, \lambda_n$,

$$|\lambda_m|^2 \leqq \sum_{i=1}^{n} |\lambda_i|^2 \leqq \sum_{j=1}^{n} \sum_{k=1}^{n} |a_{jk}|^2 \quad \textbf{(Schur's inequality)}. \tag{4}$$

In (4) *the second equality sign holds if and only if* $\mathbf{A}$ *is such that*

$$\overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A} = \mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}}. \tag{5}$$

Matrices that satisfy (5) are called **normal matrices**. It is not difficult to see that Hermitian, skew-Hermitian, and unitary matrices are normal, and so are real symmetric, skew-symmetric, and orthogonal matrices.

EXAMPLE 3 **Bounds for Eigenvalues Obtained from Schur's Inequality**

For the matrix

$$\mathbf{A} = \begin{bmatrix} 26 & -2 & 2 \\ 2 & 21 & 4 \\ 4 & 2 & 28 \end{bmatrix}$$

we obtain from Schur's inequality $|\lambda| \leqq \sqrt{1949} = 44.1475$. You may verify that the eigenvalues are 30, 25, and 20. Thus $30^2 + 25^2 + 20^2 = 1925 < 1949$; in fact, $\mathbf{A}$ is not normal. ■

The preceding theorems are valid for every real or complex square matrix. Other theorems hold for special classes of matrices only. Famous is the following one, which has various applications, for instance, in economics.

THEOREM 5

Perron's Theorem[8]

Let $\mathbf{A}$ *be a real* $n \times n$ *matrix whose entries are all positive. Then* $\mathbf{A}$ *has a positive real eigenvalue* $\lambda = \rho$ *of multiplicity* 1. *The corresponding eigenvector can be chosen with all components positive.* (*The other eigenvalues are less than* ρ *in absolute value.*)

---

[7] ISSAI SCHUR (1875–1941), German mathematician, also known by his important work in group theory.

[8] OSKAR PERRON (1880–1975) and GEORG FROBENIUS (1849–1917), German mathematicians, known for their work in potential theory, ODEs (Sec. 5.4), and group theory.

For a proof see Ref. [B3], vol. II, pp. 53–62. The theorem also holds for matrices with *nonnegative* real entries ("**Perron–Frobenius Theorem**"[8]) provided **A** is **irreducible**, that is, it cannot be brought to the following form by interchanging rows and columns; here **B** and **F** are square and **0** is a zero matrix.

$$\begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{0} & \mathbf{F} \end{bmatrix}$$

Perron's theorem has various applications, for instance, in economics. It is interesting that one can obtain from it a theorem that gives a numeric algorithm:

THEOREM 6

Collatz Inclusion Theorem[9]

Let $\mathbf{A} = [a_{jk}]$ *be a real* $n \times n$ *matrix whose elements are all positive. Let* $\mathbf{x}$ *be any real vector whose components* $x_1, \cdots, x_n$ *are positive, and let* $y_1, \cdots, y_n$ *be the components of the vector* $\mathbf{y} = \mathbf{Ax}$. *Then the closed interval on the real axis bounded by the smallest and the largest of the n quotients* $q_j = y_j/x_j$ *contains at least one eigenvalue of* $\mathbf{A}$.

PROOF We have $\mathbf{Ax} = \mathbf{y}$ or

$$\mathbf{y} - \mathbf{Ax} = \mathbf{0}. \tag{6}$$

The transpose $\mathbf{A}^\mathsf{T}$ satisfies the conditions of Theorem 5. Hence $\mathbf{A}^\mathsf{T}$ has a positive eigenvalue λ and, corresponding to this eigenvalue, an eigenvector $\mathbf{u}$ whose components u_j are all positive. Thus $\mathbf{A}^\mathsf{T}\mathbf{u} = \lambda\mathbf{u}$ and by taking the transpose we obtain $\mathbf{u}^\mathsf{T}\mathbf{A} = \lambda\mathbf{u}^\mathsf{T}$. From this and (6) we have

$$\mathbf{u}^\mathsf{T}(\mathbf{y} - \mathbf{Ax}) = \mathbf{u}^\mathsf{T}\mathbf{y} - \mathbf{u}^\mathsf{T}\mathbf{Ax} = \mathbf{u}^\mathsf{T}\mathbf{y} - \lambda\mathbf{u}^\mathsf{T}\mathbf{x} = \mathbf{u}^\mathsf{T}(\mathbf{y} - \lambda\mathbf{x}) = 0$$

or written out

$$\sum_{j=1}^{n} u_j(y_j - \lambda x_j) = 0.$$

Since all the components u_j are positive, it follows that

$$\begin{aligned} &y_j - \lambda x_j \geqq 0, \quad \text{that is,} \quad q_j \geqq \lambda \quad \text{for at least one } j, \quad \text{and} \\ &y_j - \lambda x_j \leqq 0, \quad \text{that is,} \quad q_j \leqq \lambda \quad \text{for at least one } j. \end{aligned} \tag{7}$$

Since $\mathbf{A}$ and $\mathbf{A}^\mathsf{T}$ have the same eigenvalues, λ is an eigenvalue of $\mathbf{A}$, and from (7) the statement of the theorem follows. ∎

[9]LOTHAR COLLATZ (1910–1990), German mathematician known for his work in numerics.

EXAMPLE 4 **Bounds for Eigenvalues from Collatz's Theorem. Iteration**

For a given matrix $\mathbf{A}$ with positive entries we choose an $\mathbf{x} = \mathbf{x}_0$ and **iterate**, that is, we compute $\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0$, $\mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \cdots, \mathbf{x}_{20} = \mathbf{A}\mathbf{x}_{19}$. In each step, taking $\mathbf{x} = \mathbf{x}_j$ and $\mathbf{y} = \mathbf{A}\mathbf{x}_j = \mathbf{x}_{j+1}$ we compute an inclusion interval by Collatz's theorem. This gives (6S)

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \mathbf{x}_1 = \begin{bmatrix} 0.73 \\ 0.50 \\ 0.82 \end{bmatrix}, \mathbf{x}_2 = \begin{bmatrix} 0.5481 \\ 0.3186 \\ 0.5886 \end{bmatrix},$$

$$\cdots, \mathbf{x}_{19} = \begin{bmatrix} 0.00216309 \\ 0.00108155 \\ 0.00216309 \end{bmatrix}, \mathbf{x}_{20} = \begin{bmatrix} 0.00155743 \\ 0.000778713 \\ 0.00155743 \end{bmatrix}$$

and the intervals $0.5 \leqq \lambda \leqq 0.82$, $0.3186/0.50 = 0.6372 \leqq \lambda \leqq 0.5481/0.73 = 0.750822$, etc. These intervals have length

j	1	2	3	10	15	20
Length	0.32	0.113622	0.0539835	0.0004217	0.0000132	0.0000004

Using the characteristic polynomial, you may verify that the eigenvalues of $\mathbf{A}$ are 0.72, 0.36, 0.09, so that those intervals include the largest eigenvalue, 0.72. Their lengths decreased with j, so that the iteration was worthwhile. The reason will appear in the next section, where we discuss an iteration method for eigenvalues. ■

PROBLEM SET 20.7

1–6 GERSCHGORIN DISKS

Find and sketch disks or intervals that contain the eigenvalues. If you have a CAS, find the spectrum and compare.

1. $\begin{bmatrix} 5 & 2 & 4 \\ -2 & 0 & 2 \\ 2 & 4 & 7 \end{bmatrix}$

2. $\begin{bmatrix} 5 & 10^{-2} & 10^{-2} \\ 10^{-2} & 8 & 10^{-2} \\ 10^{-2} & 10^{-2} & 9 \end{bmatrix}$

3. $\begin{bmatrix} 0 & 0.4 & -0.1 \\ -0.4 & 0 & 0.3 \\ 0.1 & -0.3 & 0 \end{bmatrix}$

4. $\begin{bmatrix} 1 & 0 & 1 \\ 0 & 4 & 3 \\ 1 & 3 & 12 \end{bmatrix}$

5. $\begin{bmatrix} 2 & i & 1+i \\ -i & 3 & 0 \\ 1-i & 0 & 8 \end{bmatrix}$

6. $\begin{bmatrix} 10 & 0.1 & -0.2 \\ 0.1 & 6 & 0 \\ -0.2 & 0 & 3 \end{bmatrix}$

7. Similarity. In Prob. 2, find $\mathbf{T}^{-1}\mathbf{A}\mathbf{T}$ such that the radius of the Gerschgorin circle with center 5 is reduced by a factor 1/100.

8. By what integer factor can you at most reduce the Gerschgorin circle with center 3 in Prob. 6?

9. If a symmetric $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ has been diagonalized except for small off-diagonal entries of size 10^{-5}, what can you say about the eigenvalues?

10. Optimality of Gerschgorin disks. Illustrate with a 2×2 matrix that an eigenvalue may very well lie on a Gerschgorin circle, so that Gerschgorin disks can generally not be replaced with smaller disks without losing the inclusion property.

11. Spectral radius $\rho(\mathbf{A})$. Using Theorem 1, show that $\rho(\mathbf{A})$ cannot be greater than the row sum norm of $\mathbf{A}$.

12–16 SPECTRAL RADIUS

Use (4) to obtain an upper bound for the spectral radius:

12. In Prob. 4

13. In Prob. 1

14. In Prob. 6

15. In Prob. 3

16. In Prob. 5

17. Verify that the matrix in Prob. 5 is normal.

18. Normal matrices. Show that Hermitian, skew-Hermitian, and unitary matrices (hence real symmetric, skew-symmetric, and orthogonal matrices) are normal. Why is this of practical interest?

19. Prove Theorem 3 by using Theorem 1.

20. Extended Gerschgorin theorem. Prove Theorem 2. *Hint.* Let $\mathbf{A} = \mathbf{B} + \mathbf{C}$, $\mathbf{B} = \text{diag}\,(a_{jj})$, $\mathbf{A}_t = \mathbf{B} + t\mathbf{C}$, and let t increase continuously from 0 to 1.

20.8 Power Method for Eigenvalues

A simple standard procedure for computing approximate values of the eigenvalues of an $n \times n$ matrix $\mathbf{A} = [a_{jk}]$ is the **power method**. In this method we start from any vector $\mathbf{x}_0$ ($\neq \mathbf{0}$) with n components and compute successively

$$\mathbf{x}_1 = \mathbf{A}\mathbf{x}_0, \qquad \mathbf{x}_2 = \mathbf{A}\mathbf{x}_1, \qquad \cdots, \qquad \mathbf{x}_s = \mathbf{A}\mathbf{x}_{s-1}.$$

For simplifying notation, we denote $\mathbf{x}_{s-1}$ by $\mathbf{x}$ and $\mathbf{x}_s$ by $\mathbf{y}$, so that $\mathbf{y} = \mathbf{A}\mathbf{x}$.

The method applies to any $n \times n$ matrix $\mathbf{A}$ that has a **dominant eigenvalue** (a λ such that $|\lambda|$ is greater than the absolute values of the other eigenvalues). If $\mathbf{A}$ is *symmetric*, it also gives the error bound (2), in addition to the approximation (1).

THEOREM 1

Power Method, Error Bounds

Let $\mathbf{A}$ *be an* $n \times n$ *real symmetric matrix. Let* $\mathbf{x}$ ($\neq \mathbf{0}$) *be any real vector with* n *components. Furthermore, let*

$$\mathbf{y} = \mathbf{A}\mathbf{x}, \qquad m_0 = \mathbf{x}^\mathsf{T}\mathbf{x}, \qquad m_1 = \mathbf{x}^\mathsf{T}\mathbf{y}, \qquad m_2 = \mathbf{y}^\mathsf{T}\mathbf{y}.$$

Then the quotient

$$q = \frac{m_1}{m_0} \qquad \textbf{(1)} \quad \textbf{(Rayleigh[10] quotient)}$$

is an approximation for an eigenvalue λ *of* $\mathbf{A}$ (usually that which is greatest in absolute value, but no general statements are possible).

Furthermore, if we set $q = \lambda - \epsilon$, *so that* ϵ *is the error of* q, *then*

$$|\epsilon| \leqq \delta = \sqrt{\frac{m_2}{m_0} - q^2}. \qquad \textbf{(2)}$$

PROOF δ^2 denotes the radicand in (2). Since $m_1 = qm_0$ by (1), we have

$$(\mathbf{y} - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = m_2 - 2qm_1 + q^2 m_0 = m_2 - q^2 m_0 = \delta^2 m_0. \qquad (3)$$

Since $\mathbf{A}$ is real symmetric, it has an orthogonal set of n real unit eigenvectors $\mathbf{z}_1, \cdots, \mathbf{z}_n$ corresponding to the eigenvalues $\lambda_1, \cdots, \lambda_n$, respectively (some of which may be equal). (Proof in Ref. [B3], vol. 1, pp. 270–272, listed in App. 1.) Then $\mathbf{x}$ has a representation of the form

$$\mathbf{x} = a_1\mathbf{z}_1 + \cdots + a_n\mathbf{z}_n.$$

[10]LORD RAYLEIGH (JOHN WILLIAM STRUTT) (1842–1919), great English physicist and mathematician, professor at Cambridge and London, known for his important contributions to various branches of applied mathematics and theoretical physics, in particular, the theory of waves, elasticity, and hydrodynamics. In 1904 he received a Nobel Prize in physics.

Now $\mathbf{A}\mathbf{z}_1 = \lambda_1\mathbf{z}_1$, etc., and we obtain

$$\mathbf{y} = \mathbf{A}\mathbf{x} = a_1\lambda_1\mathbf{z}_1 + \cdots + a_n\lambda_n\mathbf{z}_n$$

and, since the $\mathbf{z}_j$ are orthogonal unit vectors,

$$m_0 = \mathbf{x}^\mathsf{T}\mathbf{x} = a_1^2 + \cdots + a_n^2. \tag{4}$$

It follows that in (3),

$$\mathbf{y} - q\mathbf{x} = a_1(\lambda_1 - q)\mathbf{z}_1 + \cdots + a_n(\lambda_n - q)\mathbf{z}_n.$$

Since the $\mathbf{z}_j$ are orthogonal unit vectors, we thus obtain from (3)

$$\delta^2 m_0 = (y - q\mathbf{x})^\mathsf{T}(\mathbf{y} - q\mathbf{x}) = a_1^2(\lambda_1 - q)^2 + \cdots + a_n^2(\lambda_n - q)^2. \tag{5}$$

Now let λ_c be an eigenvalue of $\mathbf{A}$ to which q is closest, where c suggests "closest." Then $(\lambda_c - q)^2 \leqq (\lambda_j - q)^2$ for $j = 1, \cdots, n$. From this and (5) we obtain the inequality

$$\delta^2 m_0 \geqq (\lambda_c - q)^2(a_1^2 + \cdots + a_n^2) = (\lambda_c - q)^2 m_0.$$

Dividing by m_0, taking square roots, and recalling the meaning of δ^2 gives

$$\delta = \sqrt{\frac{m_2}{m_0} - q^2} \geqq |\lambda_c - q|.$$

This shows that δ is a bound for the error ϵ of the approximation q of an eigenvalue of $\mathbf{A}$ and completes the proof. ■

The main advantage of the method is its simplicity. And it can handle *sparse matrices* too large to store as a full square array. Its disadvantage is its possibly slow convergence. From the proof of Theorem 1 we see that the speed of convergence depends on the ratio of the dominant eigenvalue to the next in absolute value (2:1 in Example 1, below).

If we want a convergent sequence of **eigenvectors,** then at the beginning of each step we **scale** the vector, say, by dividing its components by an absolutely largest one, as in Example 1, as follows.

EXAMPLE 1 **Application of Theorem 1. Scaling**

For the symmetric matrix $\mathbf{A}$ in Example 4, Sec. 20.7, and $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ we obtain from (1) and (2) and the indicated scaling

$$\mathbf{A} = \begin{bmatrix} 0.49 & 0.02 & 0.22 \\ 0.02 & 0.28 & 0.20 \\ 0.22 & 0.20 & 0.40 \end{bmatrix}, \quad \mathbf{x}_0 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \mathbf{x}_1 = \begin{bmatrix} 0.890244 \\ 0.609756 \\ 1 \end{bmatrix}, \quad \mathbf{x}_2 = \begin{bmatrix} 0.931193 \\ 0.541284 \\ 1 \end{bmatrix}$$

$$\mathbf{x}_5 = \begin{bmatrix} 0.990663 \\ 0.504682 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{10} = \begin{bmatrix} 0.999707 \\ 0.500146 \\ 1 \end{bmatrix}, \quad \mathbf{x}_{15} = \begin{bmatrix} 0.999991 \\ 0.500005 \\ 1 \end{bmatrix}.$$

Here $\mathbf{Ax}_0 = [0.73 \quad 0.5 \quad 0.82]^\mathsf{T}$, scaled to $\mathbf{x}_1 = [0.73/0.82 \quad 0.5/0.82 \quad 1]^\mathsf{T}$, etc. The dominant eigenvalue is 0.72, an eigenvector $[1 \quad 0.5 \quad 1]^\mathsf{T}$. The corresponding q and δ are computed each time before the next scaling. Thus in the first step,

$$q = \frac{m_1}{m_0} = \frac{\mathbf{x}_0^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} = \frac{2.05}{3} = 0.683333$$

$$\delta = \left(\frac{m_2}{m_0} - q^2\right)^{1/2} = \left(\frac{(\mathbf{Ax}_0)^\mathsf{T}\mathbf{Ax}_0}{\mathbf{x}_0^\mathsf{T}\mathbf{x}_0} - q^2\right)^{1/2} = \left(\frac{1.4553}{3} - q^2\right)^{1/2} = 0.134743.$$

This gives the following values of q, δ, and the error $\epsilon = 0.72 - q$ (calculations with 10D, rounded to 6D):

j	1	2	5	10
q	0.683333	0.716048	0.719944	0.720000
δ	0.134743	0.038887	0.004499	0.000141
ϵ	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$

The error bounds are much larger than the actual errors. This is typical, although the bounds cannot be improved; that is, for special symmetric matrices they agree with the errors.

Our present results are somewhat better than those of Collatz's method in Example 4 of Sec. 20.7, at the expense of more operations. ■

Spectral shift, the transition from $\mathbf{A}$ to $\mathbf{A} - k\mathbf{I}$, shifts every eigenvalue by $-k$. Although finding a good k can hardly be made automatic, it may be helped by some other method or small preliminary computational experiments. In Example 1, Gerschgorin's theorem gives $-0.02 \leqq \lambda \leqq 0.82$ for the whole spectrum (verify!). Shifting by -0.4 might be too much (then $-0.42 \leqq \lambda \leqq 0.42$), so let us try -0.2.

EXAMPLE 2 Power Method with Spectral Shift

For $\mathbf{A} - 0.2\mathbf{I}$ with $\mathbf{A}$ as in Example 1 we obtain the following substantial improvements (where the index 1 refers to Example 1 and the index 2 to the present example).

j	1	2	5	10
δ_1	0.134743	0.038887	0.004499	0.000141
δ_2	0.134743	0.034474	0.000693	$1.8 \cdot 10^{-6}$
ϵ_1	0.036667	0.003952	0.000056	$5 \cdot 10^{-8}$
ϵ_2	0.036667	0.002477	$1.3 \cdot 10^{-6}$	$9 \cdot 10^{-12}$

■

PROBLEM SET 20.8

1–4 POWER METHOD WITHOUT SCALING

Apply the power method without scaling (3 steps), using $\mathbf{x}_0 = [1, \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1]^\mathsf{T}$. Give Rayleigh quotients and error bounds. Show the details of your work.

1. $\begin{bmatrix} 9 & 4 \\ 4 & 3 \end{bmatrix}$

2. $\begin{bmatrix} 7 & -3 \\ -3 & -1 \end{bmatrix}$

3. $\begin{bmatrix} 2 & -1 & 1 \\ -1 & 3 & 2 \\ 1 & 2 & 3 \end{bmatrix}$

4. $\begin{bmatrix} 3.6 & -1.8 & 1.8 \\ -1.8 & 2.8 & -2.6 \\ 1.8 & -2.6 & 2.8 \end{bmatrix}$

5–8 POWER METHOD WITH SCALING

Apply the power method (3 steps) with scaling, using $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$ or $[1 \quad 1 \quad 1 \quad 1]^\mathsf{T}$, as applicable. Give

Rayleigh quotients and error bounds. Show the details of your work.

5. The matrix in Prob. 3

6. $\begin{bmatrix} 4 & 2 & 3 \\ 2 & 7 & 6 \\ 3 & 6 & 4 \end{bmatrix}$

7. $\begin{bmatrix} 5 & 1 & 0 & 0 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 3 & 1 \\ 0 & 0 & 1 & 5 \end{bmatrix}$

8. $\begin{bmatrix} 2 & 4 & 0 & 1 \\ 4 & 1 & 2 & 8 \\ 0 & 2 & 5 & 2 \\ 1 & 8 & 2 & 0 \end{bmatrix}$

9. Prove that if $\mathbf{x}$ is an eigenvector, then $\delta = 0$ in (2). Give two examples.

10. Rayleigh quotient. Why does q generally approximate the eigenvalue of greatest absolute value? When will q be a good approximation?

11. Spectral shift, smallest eigenvalue. In Prob. 3 set $\mathbf{B} = \mathbf{A} - 3\mathbf{I}$ (as perhaps suggested by the diagonal entries) and see whether you may get a sequence of q's converging to an eigenvalue of $\mathbf{A}$ that *is smallest* (not largest) in absolute value. Use $\mathbf{x}_0 = [1 \quad 1 \quad 1]^\mathsf{T}$. Do 8 steps. Verify that $\mathbf{A}$ has the spectrum $\{0, 3, 5\}$.

12. CAS EXPERIMENT. Power Method with Scaling. Shifting. (a) Write a program for $n \times n$ matrices that prints every step. Apply it to the (nonsymmetric!) matrix (20 steps), starting from $[1 \quad 1 \quad 1]^\mathsf{T}$.

$$\mathbf{A} = \begin{bmatrix} 15 & 12 & 3 \\ 18 & 44 & 18 \\ -19 & -36 & -7 \end{bmatrix}.$$

(b) Experiment in (a) with shifting. Which shift do you find optimal?

(c) Write a program as in (a) but for symmetric matrices that prints vectors, scaled vectors, q, and δ. Apply it to the matrix in Prob. 8.

(d). Optimality of δ. Consider $\mathbf{A} = \begin{bmatrix} 0.6 & 0.8 \\ 0.8 & -0.6 \end{bmatrix}$ and take $\mathbf{x}_0 = \begin{bmatrix} 3 \\ -1 \end{bmatrix}$. Show that $q = 0, \delta = 1$ for all steps and the eigenvalues are ± 1, so that the interval $[q - \delta, q + \delta]$ cannot be shortened (by omitting ± 1) without losing the inclusion property. Experiment with other $\mathbf{x}_0$'s.

(e) Find a (nonsymmetric) matrix for which δ in (2) is no longer an error bound.

(f) Experiment systematically with speed of convergence by choosing matrices with the second greatest eigenvalue (i) almost equal to the greatest, (ii) somewhat different, (iii) much different.

20.9 Tridiagonalization and QR-Factorization

We consider the problem of computing *all* the eigenvalues of a ***real symmetric*** matrix $\mathbf{A} = [a_{jk}]$, discussing a method widely used in practice. In the ***first stage*** we reduce the given matrix stepwise to a **tridiagonal matrix**, that is, a matrix having all its nonzero entries on the main diagonal and in the positions immediately adjacent to the main diagonal (such as $\mathbf{A}_3$ in Fig. 450, Third Step). This reduction was invented by A. S. Householder[11] (*J. Assn. Comput. Machinery* **5** (1958), 335–342). See also Ref. [E29] in App. 1.

This Householder tridiagonalization will simplify the matrix without changing its eigenvalues. The latter will then be determined (approximately) by factoring the tridiagonalized matrix, as discussed later in this section.

[11] ALSTON SCOTT HOUSEHOLDER (1904–1993), American mathematician, known for his work in numerical analysis and mathematical biology. He was head of the mathematics division at Oakridge National Laboratory and later professor at the University of Tennessee. He was both president of ACM (Association for Computing Machinery) 1954–1956 and SIAM (Society for Industrial and Applied Mathematics) 1963–1964.

Householder's Tridiagonalization Method[11]

An $n \times n$ real symmetric matrix $\mathbf{A} = [a_{jk}]$ being given, we reduce it by $n - 2$ successive similarity transformations (see Sec. 20.6) involving matrices $\mathbf{P}_1, \cdots, \mathbf{P}_{n-2}$ to tridiagonal form. These matrices are orthogonal and symmetric. Thus $\mathbf{P}_1^{-1} = \mathbf{P}_1^{\mathsf{T}} = \mathbf{P}_1$ and similarly for the others. These transformations produce, from the given $\mathbf{A}_0 = \mathbf{A} = [a_{jk}]$, the matrices $\mathbf{A}_1 = [a_{jk}^{(1)}]$, $\mathbf{A}_2 = [a_{jk}^{(2)}], \cdots, \mathbf{A}_{n-2} = [a_{jk}^{(n-2)}]$ in the form

$$
\begin{aligned}
\mathbf{A}_1 &= \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 \\
\mathbf{A}_2 &= \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 \\
&\cdots\cdots\cdots\cdots \\
\mathbf{B} = \mathbf{A}_{n-2} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2}.
\end{aligned}
\tag{1}
$$

The transformations (1) create the necessary zeros, in the first step in Row 1 and Column 1, in the second step in Row 2 and Column 2, etc., as Fig. 450 illustrates for a 5×5 matrix. $\mathbf{B}$ is tridiagonal.

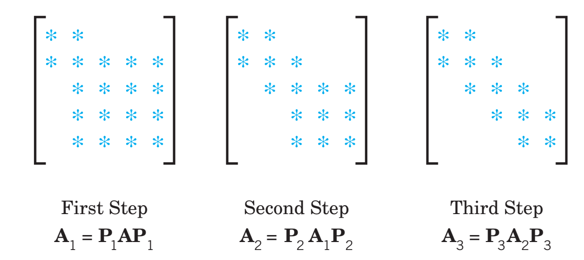


Fig. 450. Householder's method for a 5×5 matrix. Positions left blank are zeros created by the method.

How do we determine $\mathbf{P}_1, \mathbf{P}_2, \cdots, \mathbf{P}_{n-2}$? Now, all these $\mathbf{P}_r$ are of the form

$$\mathbf{P}_r = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} \qquad (r = 1, \cdots, n-2) \tag{2}$$

where $\mathbf{I}$ is the $n \times n$ unit matrix and $\mathbf{v}_r = [v_{jr}]$ is a unit vector with its first r components 0; thus

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ * \\ * \\ \vdots \\ * \end{bmatrix}, \quad \mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ * \\ \vdots \\ * \end{bmatrix}, \quad \cdots, \quad \mathbf{v}_{n-2} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ * \\ * \end{bmatrix} \tag{3}$$

where the asterisks denote the other components (which will be nonzero in general).

Step 1. $\mathbf{v}_1$ has the components

$$\text{(4)}\qquad \begin{aligned} &v_{11} = 0 \\ \text{(a)}\quad &v_{21} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{21}|}{S_1}\right)} \\ \text{(b)}\quad &v_{j1} = \frac{a_{j1}\,\text{sgn}\,a_{21}}{2v_{21}S_1} \qquad j = 3, 4, \cdots, n \\ &\text{where} \\ \text{(c)}\quad &S_1 = \sqrt{a_{21}^2 + a_{31}^2 + \cdots + a_{n1}^2} \end{aligned}$$

where $S_1 > 0$, and $\text{sgn}\,a_{21} = +1$ if $a_{21} \geqq 0$ and $\text{sgn}\,a_{21} = -1$ if $a_{21} < 0$. With this we compute $\mathbf{P}_1$ by (2) and then $\mathbf{A}_1$ by (1). This was the first step.

Step 2. We compute $\mathbf{v}_2$ by (4) with all subscripts increased by 1 and the a_{jk} replaced by $a_{jk}^{(1)}$, the entries of $\mathbf{A}_1$ just computed. Thus [see also (3)]

$$\text{(4*)}\qquad \begin{aligned} &v_{12} = v_{22} = 0 \\ &v_{32} = \sqrt{\frac{1}{2}\left(1 + \frac{|a_{32}^{(1)}|}{S_2}\right)} \\ &v_{j2} = \frac{a_{j2}^{(1)}\,\text{sgn}\,a_{32}^{(1)}}{2v_{32}S_2} \qquad j = 4, 5, \cdots, n \end{aligned}$$

where

$$S_2 = \sqrt{a_{32}^{(1)2} + a_{42}^{(1)2} + \cdots + a_{n2}^{(1)2}}.$$

With this we compute $\mathbf{P}_2$ by (2) and then $\mathbf{A}_2$ by (1).

Step 3. We compute $\mathbf{v}_3$ by (4*) with all subscripts increased by 1 and the $a_{jk}^{(1)}$ replaced by the entries $a_{jk}^{(2)}$ of $\mathbf{A}_2$, and so on.

EXAMPLE 1 Householder Tridiagonalization

Tridiagonalize the real symmetric matrix

$$\mathbf{A} = \mathbf{A}_0 = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. ***Step 1.*** We compute $S_1^2 = 4^2 + 1^2 + 1^2 = 18$ from (4c). Since $a_{21} = 4 > 0$, we have $\text{sgn}\,a_{21} = +1$ in (4b) and get from (4) by straightforward computation

$$\mathbf{v}_1 = \begin{bmatrix} 0 \\ v_{21} \\ v_{31} \\ v_{41} \end{bmatrix} = \begin{bmatrix} 0 \\ 0.98559856 \\ 0.11957316 \\ 0.11957316 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & -0.94280904 & -0.23570227 & -0.23570227 \\ 0 & -0.23570227 & 0.97140452 & -0.02859548 \\ 0 & -0.23570227 & -0.02859548 & 0.97140452 \end{bmatrix}.$$

From the first line in (1) we now get

$$\mathbf{A}_1 = \mathbf{P}_1\mathbf{A}_0\mathbf{P}_1 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & -1 & -1 \\ 0 & -1 & \frac{9}{2} & \frac{3}{2} \\ 0 & -1 & \frac{3}{2} & \frac{9}{2} \end{bmatrix}.$$

Step 2. From (4*) we compute $S_2^2 = 2$ and

$$\mathbf{v}_2 = \begin{bmatrix} 0 \\ 0 \\ v_{32} \\ v_{42} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0.92387953 \\ 0.38268343 \end{bmatrix}.$$

From this and (2),

$$\mathbf{P}_2 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \\ 0 & 0 & -1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix}.$$

The second line in (1) now gives

$$\mathbf{B}_2 = \mathbf{A}_2 = \mathbf{P}_2\mathbf{A}_1\mathbf{P}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

This matrix **B** is tridiagonal. Since our given matrix has order $n = 4$, we needed $n - 2 = 2$ steps to accomplish this reduction, as claimed. (Do you see that we got more zeros than we can expect in general?)

B is similar to **A**, as we now show in general. This is essential because **B** thus has the same spectrum as **A**, by Theorem 2 in Sec. 20.6. ■

B *Similar to* A. We assert that **B** in (1) is similar to $\mathbf{A} = \mathbf{A}_0$. The matrix $\mathbf{P}_r$ is symmetric; indeed,

$$\mathbf{P}_r^{\mathsf{T}} = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I}^{\mathsf{T}} - 2(\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^{\mathsf{T}} = \mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} = \mathbf{P}_r$$

Also, $\mathbf{P}_r$ is orthogonal because $\mathbf{v}_r$ is a unit vector, so that $\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r = 1$ and thus

$$\mathbf{P}_r\mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r^2 = (\mathbf{I} - 2\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}})^2 = \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}}$$
$$= \mathbf{I} - 4\mathbf{v}_r\mathbf{v}_r^{\mathsf{T}} + 4\mathbf{v}_r(\mathbf{v}_r^{\mathsf{T}}\mathbf{v}_r)\mathbf{v}_r^{\mathsf{T}} = \mathbf{I}.$$

Hence $\mathbf{P}_r^{-1} = \mathbf{P}_r^{\mathsf{T}} = \mathbf{P}_r$ and from (1) we now obtain

$$\begin{aligned} \mathbf{B} &= \mathbf{P}_{n-2}\mathbf{A}_{n-3}\mathbf{P}_{n-2} = \cdots \\ \cdots &= \mathbf{P}_{n-2}\mathbf{P}_{n-3}\cdots\mathbf{P}_1\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}_{n-2}^{-1}\mathbf{P}_{n-3}^{-1}\cdots\mathbf{P}_1^{-1}\mathbf{A}\mathbf{P}_1\cdots\mathbf{P}_{n-3}\mathbf{P}_{n-2} \\ &= \mathbf{P}^{-1}\mathbf{A}\mathbf{P} \end{aligned}$$

where $\mathbf{P} = \mathbf{P}_1\mathbf{P}_2\cdots\mathbf{P}_{n-2}$. This proves our assertion. ■

QR-Factorization Method

In 1958 H. Rutishauser[12] of Switzerland proposed the idea of using the LU-factorization (Sec. 20.2; he called it LR-factorization) in solving eigenvalue problems. An improved version of Rutishauser's method (avoiding breakdown if certain submatrices become singular, etc.; see Ref. [E29]) is the QR-method, independently proposed by the American J. G. F. Francis (*Computer J.* **4** (1961–62), 265–271, 332–345) and the Russian V. N. Kublanovskaya (*Zhurnal Vych. Mat. i Mat. Fiz.* **1** (1961), 555–570). The QR-method uses the factorization **QR** with orthogonal **Q** and upper triangular **R.** We discuss the **QR**-method for a real ***symmetric*** matrix. (For extensions to general matrices see Ref. [E29] in App. 1.)

In this method we first transform a given real symmetric $n \times n$ matrix $\mathbf{A}$ into a tridiagonal matrix $\mathbf{B}_0 = \mathbf{B}$ by Householder's method. This creates many zeros and thus reduces the amount of further work. Then we compute $\mathbf{B}_1, \mathbf{B}_2, \cdots$ stepwise according to the following iteration method.

Step 1. Factor $\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0$ with orthogonal $\mathbf{Q}_0$ and upper triangular $\mathbf{R}_0$. Then compute $\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0$.

Step 2. Factor $\mathbf{B}_1 = \mathbf{Q}_1\mathbf{R}_1$. Then compute $\mathbf{B}_2 = \mathbf{R}_1\mathbf{Q}_1$.

General Step s + 1.

$$\text{(5)} \qquad \begin{aligned} &\textbf{(a)} \quad \text{Factor } \mathbf{B}_s = \mathbf{Q}_s\mathbf{R}_s. \\ &\textbf{(b)} \quad \text{Compute } \mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s. \end{aligned}$$

Here $\mathbf{Q}_s$ is orthogonal and $\mathbf{R}_s$ upper triangular. The factorization (5a) will be explained below.

$\mathbf{B}_{s+1}$ Similar to B. Convergence to a Diagonal Matrix. From (5a) we have $\mathbf{R}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s$. Substitution into (5b) gives

$$\mathbf{B}_{s+1} = \mathbf{R}_s\mathbf{Q}_s = \mathbf{Q}_s^{-1}\mathbf{B}_s\mathbf{Q}_s. \tag{6}$$

[12]HEINZ RUTISHAUSER (1918–1970). Swiss mathematician, professor at ETH Zurich. Known for his pioneering work in numerics and computer science.

Thus $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_s$. Hence $\mathbf{B}_{s+1}$ is similar to $\mathbf{B}_0 = \mathbf{B}$ for all s. By Theorem 2, Sec. 20.6, this implies that $\mathbf{B}_{s+1}$ has the same eigenvalues as $\mathbf{B}$.

Also, $\mathbf{B}_{s+1}$ is symmetric. This follows by induction. Indeed, $\mathbf{B}_0 = \mathbf{B}$ is symmetric. Assuming $\mathbf{B}_s$ to be symmetric, that is, $\mathbf{B}_s{}^\mathsf{T} = \mathbf{B}_s$, and using $\mathbf{Q}_s^{-1} = \mathbf{Q}_s{}^\mathsf{T}$ (since $\mathbf{Q}_s$ is orthogonal), we get from (6) the symmetry,

$$\mathbf{B}_{s+1}{}^\mathsf{T} = (\mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s)^\mathsf{T} = Q_s{}^\mathsf{T}\mathbf{B}_s{}^\mathsf{T}\mathbf{Q}_s = \mathbf{Q}_s{}^\mathsf{T}\mathbf{B}_s\mathbf{Q}_s = \mathbf{B}_{s+1}.$$

If the eigenvalues of $\mathbf{B}$ are different in absolute value, say, $|\lambda_1| > |\lambda_2| > \cdots > |\lambda_n|$, then

$$\lim_{s\to\infty} \mathbf{B}_s = \mathbf{D}$$

where $\mathbf{D}$ is diagonal, with main diagonal entries $\lambda_1, \lambda_2, \cdots, \lambda_n$. (Proof in Ref. [E29] listed in App. 1.)

How to Get the QR-Factorization, say, $\mathbf{B} = \mathbf{B}_0 = [b_{jk}] = \mathbf{Q}_0\mathbf{R}_0$. The tridiagonal matrix $\mathbf{B}$ has $n - 1$ generally nonzero entries below the main diagonal. These are $b_{21}, b_{32}, \cdots, b_{n,n-1}$. We multiply $\mathbf{B}$ from the left by a matrix $\mathbf{C}_2$ such that $\mathbf{C}_2\mathbf{B} = [b_{jk}^{(2)}]$ has $b_{21}^{(2)} = 0$. We multiply this by a matrix $\mathbf{C}_3$ such that $\mathbf{C}_3\mathbf{C}_2\mathbf{B} = [b_{jk}^{(3)}]$ has $b_{32}^{(3)} = 0$, etc. After $n - 1$ such multiplications we are left with an upper triangular matrix $\mathbf{R}_0$, namely,

$$\mathbf{C}_n\mathbf{C}_{n-1} \cdots \mathbf{C}_3\mathbf{C}_2 B_0 = \mathbf{R}_0. \tag{7}$$

These $n \times n$ matrices $\mathbf{C}_j$ are very simple. $\mathbf{C}_j$ has the 2×2 submatrix

$$\begin{bmatrix} \cos\theta_j & \sin\theta_j \\ -\sin\theta_j & \cos\theta_j \end{bmatrix} \qquad (\theta_j \text{ suitable})$$

in Rows $j - 1$ and j and Columns $j - 1$ and j; everywhere else on the main diagonal the matrix $\mathbf{C}_j$ has entries 1; and all its other entries are 0. (This submatrix is the matrix of a plane rotation through the angle θ_j; see Team Project 30, Sec. 7.2.) For instance, if $n = 4$, writing $c_j = \cos\theta_j$, $s_j = \sin\theta_j$, we have

$$\mathbf{C}_2 = \begin{bmatrix} c_2 & s_2 & 0 & 0 \\ -s_2 & c_2 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & c_3 & s_3 & 0 \\ 0 & -s_3 & c_3 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}, \quad \mathbf{C}_4 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & c_4 & s_4 \\ 0 & 0 & -s_4 & c_4 \end{bmatrix}.$$

These $\mathbf{C}_j$ are orthogonal. Hence their product in (7) is orthogonal, and so is the inverse of this product. We call this inverse $\mathbf{Q}_0$. Then from (7),

$$\mathbf{B}_0 = \mathbf{Q}_0\mathbf{R}_0 \tag{8}$$

where, with $\mathbf{C}_j^{-1} = \mathbf{C}_j{}^\mathsf{T}$,

$$\mathbf{Q}_0 = (\mathbf{C}_n\mathbf{C}_{n-1} \cdots \mathbf{C}_3\mathbf{C}_2)^{-1} = \mathbf{C}_2{}^\mathsf{T}\mathbf{C}_3{}^\mathsf{T} \cdots \mathbf{C}_{n-1}{}^\mathsf{T}\mathbf{C}_n{}^\mathsf{T}. \tag{9}$$

This is our QR-factorization of $\mathbf{B}_0$. From it we have by (5b) with $s = 0$

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{Q}_0 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}} \cdots \mathbf{C}_{n-1}{}^{\mathsf{T}}\mathbf{C}_n{}^{\mathsf{T}}. \tag{10}$$

We do not need $\mathbf{Q}_0$ explicitly, but to get $\mathbf{B}_1$ from (10), we first compute $\mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}$, then $(\mathbf{R}_0\mathbf{C}_2^{\mathsf{T}})\mathbf{C}_3^{\mathsf{T}}$, etc. Similarly in the further steps that produce $\mathbf{B}_2, \mathbf{B}_3, \cdots$.

Determination of cos θ_j and sin θ_j. We finally show how to find the angles of rotation. $\cos\theta_2$ and $\sin\theta_2$ in $\mathbf{C}_2$ must be such that $b_{21}^{(2)} = 0$ in the product

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} c_2 & s_2 & 0 & \cdots \\ -s_2 & c_2 & 0 & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} & b_{13} & \cdots \\ b_{21} & b_{22} & b_{23} & \cdots \\ \cdot & \cdot & \cdot & \cdots \\ \cdot & \cdot & \cdot & \cdots \end{bmatrix}.$$

Now $b_{21}^{(2)}$ is obtained by multiplying the second row of $\mathbf{C}_2$ by the first column of $\mathbf{B}$,

$$b_{21}^{(2)} = -s_2 b_{11} + c_2 b_{21} = -(\sin\theta_2)b_{11} + (\cos\theta_2)b_{21} = 0.$$

Hence $\tan\theta_2 = s_2/c_2 = b_{21}/b_{11}$, and

$$\begin{aligned} \cos\theta_2 &= \frac{1}{\sqrt{1 + \tan^2\theta_2}} = \frac{1}{\sqrt{1 + (b_{21}/b_{11})^2}} \\ \sin\theta_2 &= \frac{\tan\theta_2}{\sqrt{1 + \tan^2\theta_2}} = \frac{b_{21}/b_{11}}{\sqrt{1 + (b_{21}/b_{11})^2}}. \end{aligned} \tag{11}$$

Similarly for $\theta_3, \theta_4, \cdots$. The next example illustrates all this.

EXAMPLE 2 QR-Factorization Method

Compute all the eigenvalues of the matrix

$$\mathbf{A} = \begin{bmatrix} 6 & 4 & 1 & 1 \\ 4 & 6 & 1 & 1 \\ 1 & 1 & 5 & 2 \\ 1 & 1 & 2 & 5 \end{bmatrix}.$$

Solution. We first reduce $\mathbf{A}$ to tridiagonal form. Applying Householder's method, we obtain (see Example 1)

$$\mathbf{A}_2 = \begin{bmatrix} 6 & -\sqrt{18} & 0 & 0 \\ -\sqrt{18} & 7 & \sqrt{2} & 0 \\ 0 & \sqrt{2} & 6 & 0 \\ 0 & 0 & 0 & 3 \end{bmatrix}.$$

From the characteristic determinant we see that $\mathbf{A}_2$, hence $\mathbf{A}$, has the eigenvalue 3. (Can you see this directly from $\mathbf{A}_2$?) Hence it suffices to apply the QR-method to the tridiagonal 3×3 matrix

$$\mathbf{B}_0 = \mathbf{B} = \begin{bmatrix} 6 & -\sqrt{18} & 0 \\ -\sqrt{18} & 7 & \sqrt{2} \\ 0 & \sqrt{2} & 6 \end{bmatrix}.$$

Step 1. We multiply $\mathbf{B}$ from the left by

$$\mathbf{C}_2 = \begin{bmatrix} \cos\theta_2 & \sin\theta_2 & 0 \\ -\sin\theta_2 & \cos\theta_2 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and then } \mathbf{C}_2\mathbf{B} \text{ by} \quad \mathbf{C}_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_3 & \sin\theta_3 \\ 0 & -\sin\theta_3 & \cos\theta_3 \end{bmatrix}.$$

Here $(-\sin\theta_2) \cdot 6 + (\cos\theta_2)(-\sqrt{18}) = 0$ gives (11) $\cos\theta_2 = 0.81649658$ and $\sin\theta_2 = -0.57735027$. With these values we compute

$$\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.26598632 & 1.15470054 \\ 0 & 1.41421356 & 6.00000000 \end{bmatrix}.$$

In $\mathbf{C}_3$ we get from $(-\sin\theta_3) \cdot 3.26598632 + (\cos\theta_3) \cdot 1.41421356 = 0$ the values $\cos\theta_3 = 0.91766294$ and $\sin\theta_3 = 0.39735971$. This gives

$$\mathbf{R}_0 = \mathbf{C}_3\mathbf{C}_2\mathbf{B} = \begin{bmatrix} 7.34846923 & -7.50555350 & -0.81649658 \\ 0 & 3.55902608 & 3.44378413 \\ 0 & 0 & 5.04714615 \end{bmatrix}.$$

From this we compute

$$\mathbf{B}_1 = \mathbf{R}_0\mathbf{C}_2{}^{\mathsf{T}}\mathbf{C}_3{}^{\mathsf{T}} = \begin{bmatrix} 10.33333333 & -2.05480467 & 0 \\ -2.05480467 & 4.03508772 & 2.00553251 \\ 0 & 2.00553251 & 4.63157895 \end{bmatrix}$$

which is symmetric and tridiagonal. The off-diagonal entries in $\mathbf{B}_1$ are still large in absolute value. Hence we have to go on.

Step 2. We do the same computations as in the first step, with $\mathbf{B}_0 = \mathbf{B}$ replaced by $\mathbf{B}_1$ and $\mathbf{C}_2$ and $\mathbf{C}_3$ changed accordingly, the new angles being $\theta_2 = -0.196291533$ and $\theta_3 = 0.513415589$. We obtain

$$\mathbf{R}_1 = \begin{bmatrix} 10.53565375 & -2.80232241 & -0.39114588 \\ 0 & 4.08329584 & 3.98824028 \\ 0 & 0 & 3.06832668 \end{bmatrix}$$

and from this

$$\mathbf{B}_2 = \begin{bmatrix} 10.87987988 & -0.79637918 & 0 \\ -0.79637918 & 5.44738664 & 1.50702500 \\ 0 & 1.50702500 & 2.67273348 \end{bmatrix}.$$

We see that the off-diagonal entries are somewhat smaller in absolute value than those of $\mathbf{B}_1$, but still much too large for the diagonal entries to be good approximations of the eigenvalues of $\mathbf{B}$.

Further Steps. We list the main diagonal entries and the absolutely largest off-diagonal entry, which is $|b_{12}^{(j)}| = |b_{21}^{(j)}|$ in all steps. You may show that the given matrix **A** has the spectrum 11, 6, 3, 2.

Step j	$b_{11}^{(j)}$	$b_{22}^{(j)}$	$b_{33}^{(j)}$	$\max_{j \neq k} \lvert b_{jk}^{(J)} \rvert$
3	10.9668929	5.94589856	2.08720851	0.58523582
5	10.9970872	6.00181541	2.00109738	0.12065334
7	10.9997421	6.00024439	2.00001355	0.03591107
9	10.9999772	6.00002267	2.00000017	0.01068477

Looking back at our discussion, we recognize that the purpose of applying Householder's tridiagonalization before the QR-factorization method is a substantial reduction of cost in each QR-factorization, in particular if **A** is large.

Convergence acceleration and thus further reduction of cost can be achieved by a **spectral shift,** that is, by taking $\mathbf{B}_s - k_s\mathbf{I}$ instead of $\mathbf{B}_s$ with a suitable k_s. Possible choices of k_s are discussed in Ref. [E29], p. 510.

PROBLEM SET 20.9

1–5 HOUSEHOLDER TRIDIAGONALIZATION

Tridiagonalize. Show the details.

1. $\begin{bmatrix} 0.98 & 0.04 & 0.44 \\ 0.04 & 0.56 & 0.40 \\ 0.44 & 0.40 & 0.80 \end{bmatrix}$

2. $\begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

3. $\begin{bmatrix} 7 & 2 & 3 \\ 2 & 10 & 6 \\ 3 & 6 & 7 \end{bmatrix}$

4. $\begin{bmatrix} 5 & 4 & 1 & 1 \\ 4 & 5 & 1 & 1 \\ 1 & 1 & 4 & 2 \\ 1 & 1 & 2 & 4 \end{bmatrix}$

5. $\begin{bmatrix} 3 & 52 & 10 & 42 \\ 52 & 59 & 44 & 80 \\ 10 & 44 & 39 & 42 \\ 42 & 80 & 42 & 35 \end{bmatrix}$

6–9 QR-FACTORIZATION

Do three QR-steps to find approximations of the eigenvalues of:

6. The matrix in the answer to Prob. 1

7. The matrix in the answer to Prob. 3

8. $\begin{bmatrix} 14.2 & -0.1 & 0 \\ -0.1 & -6.3 & 0.2 \\ 0 & 0.2 & 2.1 \end{bmatrix}$

9. $\begin{bmatrix} 140 & 10 & 0 \\ 10 & 70 & 2 \\ 0 & 2 & -30 \end{bmatrix}$

10. CAS EXPERIMENT. QR-Method. Try to find out experimentally on what properties of a matrix the speed of decrease of off-diagonal entries in the QR-method depends. For this purpose write a program that first tridiagonalizes and then does QR-steps. Try the program out on the matrices in Probs. 1, 3, and 4. Summarize your findings in a short report.

CHAPTER 20 REVIEW QUESTIONS AND PROBLEMS

1. What are the main problem areas in numeric linear algebra?

2. When would you apply Gauss elimination and when Gauss–Seidel iteration?

3. What is pivoting? Why and how is it done?

4. What happens if you apply Gauss elimination to a system that has no solutions?

5. What is Cholesky's method? When would you apply it?

6. What do you know about the convergence of the Gauss–Seidel iteration?

7. What is ill-conditioning? What is the condition number and its significance?

8. Explain the idea of least squares approximation.

9. What are eigenvalues of a matrix? Why are they important? Give typical examples.

10. How did we use similarity transformations of matrices in designing numeric methods?

11. What is the power method for eigenvalues? What are its advantages and disadvantages?

12. State Gerschgorin's theorem from memory. Give typical applications.

13. What is tridiagonalization and QR? When would you apply it?

14–17 GAUSS ELIMINATION

Solve

14. $$\begin{aligned} 3x_2 - 6x_3 &= 0 \\ 4x_1 - x_2 + 2x_3 &= 16 \\ -5x_1 + 2x_2 - 4x_3 &= -20 \end{aligned}$$

15. $$\begin{aligned} 8x_2 - 6x_3 &= 23.6 \\ 10x_1 + 6x_2 + 2x_3 &= 68.4 \\ 12x_1 - 14x_2 + 4x_3 &= -6.2 \end{aligned}$$

16. $$\begin{aligned} 5x_1 + x_2 - 3x_3 &= 17 \\ -5x_2 + 15x_3 &= -10 \\ 2x_1 - 3x_2 + 9x_3 &= 0 \end{aligned}$$

17. $$\begin{aligned} 42x_1 + 74x_2 + 36x_3 &= 96 \\ -46x_1 - 12x_2 - 2x_3 &= 82 \\ 3x_1 + 25x_2 + 5x_3 &= 19 \end{aligned}$$

18–20 INVERSE MATRIX

Compute the inverse of:

18. $$\begin{bmatrix} 2.0 & 0.1 & 3.3 \\ 1.6 & 4.4 & 0.5 \\ 0.3 & -4.3 & 2.8 \end{bmatrix}$$

19. $$\begin{bmatrix} 15 & 20 & 10 \\ 20 & 35 & 15 \\ 10 & 15 & 90 \end{bmatrix}$$

20. $$\begin{bmatrix} 5 & 1 & 1 \\ 1 & 6 & 0 \\ 1 & 0 & 8 \end{bmatrix}$$

21–23 GAUSS–SEIDEL ITERATION

Do 3 steps without scaling, starting from $[1 \quad 1 \quad 1]^T$.

21. $$\begin{aligned} 4x_1 - x_2 \qquad &= 22.0 \\ 4x_2 - x_3 &= 13.4 \\ -x_1 \qquad + 4x_3 &= -2.4 \end{aligned}$$

22. $$\begin{aligned} 0.2x_1 + 4.0x_2 - 0.4x_3 &= 32.0 \\ 0.5x_1 - 0.2x_2 + 2.5x_3 &= -5.1 \\ 7.5x_1 + 0.1x_2 - 1.5x_3 &= -12.7 \end{aligned}$$

23. $$\begin{aligned} 10x_1 + x_2 - x_3 &= 17 \\ 2x_1 + 20x_2 + x_3 &= 28 \\ 3x_1 - x_2 + 25x_3 &= 105 \end{aligned}$$

24–26 VECTOR NORMS

Compute the ℓ_1-, ℓ_2-, and ℓ_∞-norms of the vectors.

24. $[0.2 \quad -8.1 \quad 0.4 \quad 0 \quad 0 \quad -1.3 \quad 2]^T$

25. $[8 \quad -21 \quad 13 \quad 0]^T$

26. $[0 \quad 0 \quad 0 \quad -1 \quad 0]^T$

27–30 MATRIX NORM

Compute the matrix norm corresponding to the ℓ_∞-vector norm for the coefficient matrix:

27. In Prob. 15

28. In Prob. 17

29. In Prob. 21

30. In Prob. 22

31–33 CONDITION NUMBER

Compute the condition number (corresponding to the ℓ_∞-vector norm) of the coefficient matrix:

31. In Prob. 19

32. In Prob. 18

33. In Prob. 21

34–35 FITTING BY LEAST SQUARES

Fit and graph:

34. A straight line to $(-1, 0)$, $(0, 2)$, $(1, 2)$, $(2, 3)$, $(3, 3)$

35. A quadratic parabola to the data in Prob. 34.

36–39 EIGENVALUES

Find and graph three circular disks that must contain all the eigenvalues of the matrix:

36. In Prob. 18

37. In Prob. 19

38. In Prob. 20

39. Of the coefficients in Prob. 14

40. Power method. Do 4 steps with scaling for the matrix in Prob. 19, starting for [1 1 1] and computing the Rayliegh quotients and error bounds.

SUMMARY OF CHAPTER 20
Numeric Linear Algebra

Main tasks are the numeric solution of linear systems (Secs. 20.1–20.4), curve fitting (Sec. 20.5), and eigenvalue problems (Secs. 20.6–20.9).

Linear systems Ax = b with $\mathbf{A} = [a_{jk}]$, written out

$$\text{(1)}\qquad \begin{aligned} \text{E}_1\text{:}&\quad a_{11}x_1 + \cdots + a_{1n}x_n = b_1 \\ \text{E}_2\text{:}&\quad a_{21}x_1 + \cdots + a_{2n}x_n = b_2 \\ &\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \text{E}_n\text{:}&\quad a_{n1}x_1 + \cdots + a_{nn}x_n = b_n \end{aligned}$$

can be solved by a **direct method** (one in which the number of numeric operations can be specified in advance, e.g., Gauss's elimination) or by an **indirect** or **iterative method** (in which an initial approximation is improved stepwise).

The **Gauss elimination** (Sec. 20.1) is direct, namely, a systematic elimination process that reduces (1) stepwise to triangular form. In Step 1 we eliminate x_1 from equations E_2 to E_n by subtracting $(a_{21}/a_{11})\,\text{E}_1$ from E_2, then $(a_{31}/a_{11})\,\text{E}_1$ from E_3, etc. Equation E_1 is called the **pivot equation** in this step and a_{11} the **pivot.** In Step 2 we take the new second equation as pivot equation and eliminate x_2, etc. If the triangular form is reached, we get x_n from the last equation, then x_{n-1} from the second last, etc. **Partial pivoting** (= interchange of equations) is *necessary* if candidates for pivots are zero, and *advisable* if they are small in absolute value.

Doolittle's, Crout's, and **Cholesky's methods** in Sec. 20.2 are variants of the Gauss elimination. They factor $\mathbf{A} = \mathbf{LU}$ ($\mathbf{L}$ lower triangular, $\mathbf{U}$ upper triangular) and solve $\mathbf{Ax} = \mathbf{LUx} = \mathbf{b}$ by solving $\mathbf{Ly} = \mathbf{b}$ for $\mathbf{y}$ and then $\mathbf{Ux} = \mathbf{y}$ for $\mathbf{x}$.

In the **Gauss–Seidel iteration** (Sec. 20.3) we make $a_{11} = a_{22} = \cdots = a_{nn} = 1$ (by division) and write $\mathbf{Ax} = (\mathbf{I} + \mathbf{L} + \mathbf{U})\mathbf{x} = \mathbf{b}$; thus $\mathbf{x} = \mathbf{b} - (\mathbf{L} + \mathbf{U})\mathbf{x}$, which suggests the iteration formula

$$\text{(2)}\qquad \mathbf{x}^{(m+1)} = \mathbf{b} - \mathbf{L}\mathbf{x}^{(m+1)} - \mathbf{U}\mathbf{x}^{(m)}$$

in which we always take the most recent approximate x_j's on the right. If $\|\mathbf{C}\| < 1$, where $\mathbf{C} = -(\mathbf{I} + \mathbf{L})^{-1}\mathbf{U}$, then this process converges. Here, $\|\mathbf{C}\|$ denotes any matrix norm (Sec. 20.3).

If the **condition number** $k(\mathbf{A}) = \|\mathbf{A}\|\,\|\mathbf{A}^{-1}\|$ of $\mathbf{A}$ is large, then the system $\mathbf{Ax} = \mathbf{b}$ is **ill-conditioned** (Sec. 20.4), and a small **residual** $\mathbf{r} = \mathbf{b} - \mathbf{A}\tilde{\mathbf{x}}$ does ***not*** imply that $\tilde{\mathbf{x}}$ is close to the exact solution.

The fitting of a polynomial $p(x) = b_0 + b_1x + \cdots + b_mx^m$ through given data (points in the xy-plane) $(x_1, y_1), \cdots, (x_n, y_n)$ by the method of **least squares** is discussed in Sec. 20.5 (and in statistics in Sec. 25.9). If $m = n$, the least squares polynomial will be the same as an interpolating polynomial (uniqueness).

Eigenvalues λ (values λ for which $\mathbf{Ax} = \lambda\mathbf{x}$ has a solution $\mathbf{x} \neq \mathbf{0}$, called an **eigenvector**) can be characterized by inequalities (Sec. 20.7), e.g. in **Gerschgorin's theorem,** which gives n circular disks which contain the whole spectrum (all eigenvalues) of $\mathbf{A}$, of centers a_{jj} and radii $\Sigma|a_{jk}|$ (sum over k from 1 to n, $k \neq j$).

Approximations of eigenvalues can be obtained by iteration, starting from an $\mathbf{x}_0 \neq \mathbf{0}$ and computing $\mathbf{x}_1 = \mathbf{Ax}_0$, $\mathbf{x}_2 = \mathbf{Ax}_1, \cdots, \mathbf{x}_n = \mathbf{Ax}_{n-1}$. In this **power method** (Sec. 20.8) the **Rayleigh quotient**

$$q = \frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{x}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} \qquad (\mathbf{x} = \mathbf{x}_n) \tag{3}$$

gives an approximation of an eigenvalue (usually that of the greatest absolute value) and, if $\mathbf{A}$ is symmetric, an error bound is

$$|\epsilon| \leqq \sqrt{\frac{(\mathbf{Ax})^{\mathsf{T}}\mathbf{Ax}}{\mathbf{x}^{\mathsf{T}}\mathbf{x}} - q^2}. \tag{4}$$

Convergence may be slow but can be improved by a *spectral shift.*

For determining all the eigenvalues of a symmetric matrix $\mathbf{A}$ it is best to first tridiagonalize $\mathbf{A}$ and then to apply the QR-method (Sec. 20.9), which is based on a factorization $\mathbf{A} = \mathbf{QR}$ with orthogonal $\mathbf{Q}$ and upper triangular $\mathbf{R}$ and uses similarity transformations.

CHAPTER 21

Numerics for ODEs and PDEs

Ordinary differential equations (ODEs) and partial differential equations (PDEs) play a central role in modeling problems of engineering, mathematics, physics, aeronautics, astronomy, dynamics, elasticity, biology, medicine, chemistry, environmental science, economics, and many other areas. Chapters 1–6 and 12 explained the major approaches to solving ODEs and PDEs analytically. However, in your career as an engineer, applied mathematicians, or physicist you will encounter ODEs and PDEs that *cannot* be solved by those analytic methods or whose solutions are so difficult that other approaches are needed. It is precisely in these real-world projects that numeric methods for ODEs and PDEs are used, often as part of a software package. Indeed, numeric software has become an indispensable tool for the engineer.

This chapter is evenly divided between numerics for ODEs and numerics for PDEs. We start with ODEs and discuss, in Sec. 21.1, methods for first-order ODEs. The main initial idea is that we can obtain approximations to the solution of such an ODE at points that are a distance h apart by using the first two terms of Taylor's formula from calculus. We use these approximations to construct the iteration formula for a method known as Euler's method. While this method is rather unstable and of little practical use, it serves as a pedagogical tool and a starting point toward understanding more sophisticated methods such as the Runge–Kutta method and its variant the Runga–Kutta–Fehlberg (RKF) method, which are popular and useful in practice. As is usual in mathematics, one tends to generalize mathematical ideas. The methods of Sec. 21.1 are one-step methods, that is, the current approximation uses only the approximation from the previous step. Multistep methods, such as the Adams–Bashforth methods and Adams–Moulton methods, use values computed from several previous steps. We conclude numerics for ODEs with applying Runge–Kutta–Nyström methods and other methods to higher order ODEs and systems of ODEs.

Numerics for PDEs are perhaps even more exciting and ingenious than those for ODEs. We first consider PDEs of the elliptic type (Laplace, Poisson). Again, Taylor's formula serves as a starting point and lets us replace partial derivatives by difference quotients. The end result leads to a mesh and an evaluation scheme that uses the Gauss–Seidel method (here also know as Liebmann's method). We continue with methods that use grids to solve Neuman and mixed problems (Sec. 21.5) and conclude with the important Crank–Nicholson method for parabolic PDEs in Sec. 21.6.

***Sections* 21.1 *and* 21.2 *may be studied immediately after Chap.* 1 *and Sec.* 21.3 *immediately after Chaps.* 2–4,** because these sections are independent of Chaps. 19 and 20.

***Sections* 21.4–21.7 *on PDEs may be studied immediately after Chap.* 12** if students have some knowledge of linear systems of algebraic equations.

Prerequisite: Secs. 1.1–1.5 for ODEs, Secs. 12.1–12.3, 12.5, 12.10 for PDEs.
References and Answers to Problems: App. 1 Part E (see also Parts A and C), App. 2.

21.1 Methods for First-Order ODEs

Take a look at Sec. 1.2, where we briefly introduced Euler's method with an example. *We shall develop* ***Euler's method*** *more rigorously.* Pay close attention to the derivation that uses Taylor's formula from calculus to approximate the solution to a first-order ODE at points that are a distance h apart. If you understand this approach, which is typical for numerics for ODEs, then you will understand other methods more easily.

From Chap. 1 we know that an ODE of the first order is of the form $F(x, y, y') = 0$ and can often be written in the explicit form $y' = f(x, y)$. An **initial value problem** for this equation is of the form

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

where x_0 and y_0 are given and we assume that the problem has a unique solution on some open interval $a < x < b$ containing x_0.

In this section we shall discuss methods of computing approximate numeric values of the solution $y(x)$ of (1) at the equidistant points on the x-axis

$$x_1 = x_0 + h, \qquad x_2 = x_0 + 2h, \qquad x_3 = x_0 + 3h, \qquad \cdots$$

where the **step size** h is a fixed number, for instance, 0.2 or 0.1 or 0.01, whose choice we discuss later in this section. Those methods are **step-by-step methods**, using the same formula in each step. Such formulas are suggested by the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots. \tag{2}$$

Formula (2) is the key idea that lets us develop Euler's method and its variant called—you guessed it—*improved Euler method*, also known as *Heun's method*. Let us start by deriving Euler's method.

For small h the higher powers $h^2, h^3, \cdots$ in (2) are very small. Dropping all of them gives the crude approximation

$$\begin{aligned} y(x + h) &\approx y(x) + hy'(x) \\ &= y(x) + hf(x, y) \end{aligned}$$

and the corresponding **Euler method** (or **Euler–Cauchy method**)

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots) \tag{3}$$

discussed in Sec. 1.2. Geometrically, this is an approximation of the curve of $y(x)$ by a polygon whose first side is tangent to this curve at x_0 (see Fig. 8 in Sec. 1.2).

Error of the Euler Method. Recall from calculus that Taylor's formula with remainder has the form

$$y(x + h) = y(x) + hy'(x) + \tfrac{1}{2} h^2 y''(\xi)$$

(where $x \leqq \xi \leqq x + h$). It shows that, in the Euler method, the *truncation error in each step* or **local truncation error** is proportional to h^2, written $O(h^2)$, where O suggests *order* (see also Sec. 20.1). Now, over a fixed x-interval in which we want to solve an ODE, the number of steps is proportional to $1/h$. Hence the *total error* or **global error** is proportional to $h^2(1/h) = h^1$. For this reason, the Euler method is called a **first-order method**. In addition, there are **roundoff errors** in this and other methods, which may affect the accuracy of the values $y_1, y_2, \cdots$ more and more as n increases.

Automatic Variable Step Size Selection in Modern Software. The idea of adaptive integration, as motivated and explained in Sec. 19.5, applies equally well to the numeric solution of ODEs. It now concerns automatically changing the step size h depending on the variability of $y' = f$ determined by

$$y'' = f' = f_x + f_y y' = f_x + f_y f. \tag{4*}$$

Accordingly, modern software automatically selects variable step sizes h_n so that the error of the solution will not exceed a given maximum size TOL (suggesting *tolerance*). Now for the Euler method, when the step size is $h = h_n$, the local error at x_n is about $\frac{1}{2}h_n^2|y''(\xi_n)|$. We require that this be equal to a given tolerance TOL,

$$\text{(a)} \quad \tfrac{1}{2}h_n^2|y''(\xi_n)| = \text{TOL}, \qquad \text{thus} \qquad \text{(b)} \quad h_n = \sqrt{\frac{2\ \text{TOL}}{|y''(\xi_n)|}}. \tag{4}$$

$y''(x)$ must not be zero on the interval J: $x_0 \leqq x = x_N$ on which the solution is wanted. Let K be the minimum of $|y''(x)|$ on J and assume that $K > 0$. Minimum $|y''(x)|$ corresponds to maximum $h = H = \sqrt{2\ \text{TOL}/K}$ by (4). Thus, $\sqrt{2\ \text{TOL}} = H\sqrt{K}$. We can insert this into (4b), obtaining by straightforward algebra

$$h_n = \varphi(x_n)H \qquad \text{where} \qquad \varphi(x_n) = \sqrt{\frac{K}{|y''(\xi_n)|}}. \tag{5}$$

For other methods, automatic step size selection is based on the same principle.

Improved Euler Method. Predictor, Corrector. Euler's method is generally much too inaccurate. For a large h (0.2) this is illustrated in Sec. 1.2 by the computation for

$$y' = y + x, \qquad y(0) = 0. \tag{6}$$

And for small h the computation becomes prohibitive; also, roundoff in so many steps may result in meaningless results. Clearly, methods of higher order and precision are obtained by taking more terms in (2) into account. But this involves an important practical problem. Namely, if we substitute $y' = f(x, y(x))$ into (2), we have

$$y(x + h) = y(x) + hf + \tfrac{1}{2}h^2 f' + \tfrac{1}{6}h^3 f'' + \cdots. \tag{2*}$$

Now y in f depends on x, so that we have f' as shown in (4*) and f'', f''' even much more cumbersome. The ***general strategy*** now is to avoid the computation of these derivatives and to replace it by computing f for one or several suitably chosen auxiliary values of (x, y). "Suitably" means that these values are chosen to make the order of the method as

high as possible (to have high accuracy). Let us discuss two such methods that are of practical importance, namely, the improved Euler method and the (classical) Runge–Kutta method.

In each step of the **improved Euler method** we compute *two* values, first the ***predictor***

(7a) $$y^*_{n+1} = y_n + hf(x_n, y_n),$$

which is an auxiliary value, and then the new y-value, the **corrector**

(7b) $$y_{n+1} = y_n + \tfrac{1}{2}h\,[f(x_n, y_n) + f(x_{n+1}, y^*_{n+1})].$$

Hence the improved Euler method is a predictor–corrector method: In each step we predict a value (7a) and then we correct it by (7b).

In algorithmic form, using the notations $k_1 = hf(x_n, y_n)$ in (7a) and $k_2 = hf(x_{n+1}, y^*_{n+1})$ in (7b), we can write this method as shown in Table 21.1.

Table 21.1 Improved Euler Method (Heun's Method)

ALGORITHM EULER (f, x_0, y_0, h, N)

This algorithm computes the solution of the initial value problem $y' = f(x, y)$, $y(x_0) = y_0$ at equidistant points $x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh$; here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.6).

INPUT: Initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$x_{n+1} = x_n + h$

$k_1 = hf(x_n, y_n)$

$k_2 = hf(x_{n+1}, y_n + k_1)$

$y_{n+1} = y_n + \tfrac{1}{2}(k_1 + k_2)$

OUTPUT x_{n+1}, y_{n+1}

End

Stop

End EULER

EXAMPLE 1 Improved Euler Method. Comparison with Euler Method.

Apply the improved Euler method to the initial value problem (6), choosing $h = 0.2$ as in Sec. 1.2.

Solution. For the present problem we have in Table 21.1

$$k_1 = 0.2(x_n + y_n)$$

$$k_2 = 0.2(x_n + 0.2 + y_n + 0.2(x_n + y_n))$$

$$y_{n+1} = y_n + \frac{0.2}{2}(2.2x_n + 2.2y_n + 0.2) = y_n + 0.22(x_n + y_n) + 0.02.$$

Table 21.2 shows that our present results are much more accurate than those for Euler's method in Table 21.1 but at the cost of more computations. ■

Table 21.2 Improved Euler Method for (6). Errors

n	x_n	y_n	Exact Values (4D)	Error of Improved Euler	Error of Euler
0	0.0	0.0000	0.0000	0.0000	0.000
1	0.2	0.0200	0.0214	0.0014	0.021
2	0.4	0.0884	0.0918	0.0034	0.052
3	0.6	0.2158	0.2221	0.0063	0.094
4	0.8	0.4153	0.4255	0.0102	0.152
5	1.0	0.7027	0.7183	0.0156	0.230

Error of the Improved Euler Method. *The local error is of order h^3 and the global error of order h^2, so that the method is a* **second-order method**.

PROOF Setting $\tilde{f}_n = f(x_n, y(x_n))$ and using (2*) (after (6)), we have

(8a) $$y(x_n + h) - y(x_n) = h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}'_n + \tfrac{1}{6}h^3\tilde{f}''_n + \cdots.$$

Approximating the expression in the brackets in (7b) by $\tilde{f}_n + \tilde{f}_{n+1}$ and again using the Taylor expansion, we obtain from (7b)

(8b) $$\begin{aligned} y_{n+1} - y_n &\approx \tfrac{1}{2}h\,[\tilde{f}_n + \tilde{f}_{n+1}] \\ &= \tfrac{1}{2}h\,[\tilde{f}_n + (\tilde{f}_n + h\tilde{f}'_n + \tfrac{1}{2}h^2\tilde{f}''_n + \cdots)] \\ &= h\tilde{f}_n + \tfrac{1}{2}h^2\tilde{f}'_n + \tfrac{1}{4}h^3\tilde{f}''_n + \cdots \end{aligned}$$

(where $' = d/dx_n$, etc.). Subtraction of (8b) from (8a) gives the local error

$$\frac{h^3}{6}\tilde{f}''_n - \frac{h^3}{4}\tilde{f}''_n + \cdots = -\frac{h^3}{12}\tilde{f}''_n + \cdots.$$

Since the number of steps over a fixed x-interval is proportional to $1/h$, the global error is of order $h^3/h = h^2$, so that the method is of second order. ■

Since the Euler method was an attractive pedagogical tool to teach the beginning of solving first-order ODEs numerically but had its drawbacks in terms of accuracy and could even produce wrong answers, we studied the improved Euler method and thereby introduced the idea of a predictor–corrector method. Although improved Euler is better than Euler, there are better methods that are used in industrial settings. Thus the practicing engineer has to know about the Runga–Kutta methods and its variants.

Runge–Kutta Methods (RK Methods)

A method of great practical importance and much greater accuracy than that of the improved Euler method is the *classical Runge–Kutta method of fourth order*, which we

call briefly the **Runge–Kutta method.**[1] It is shown in Table 21.3. We see that in each step we first compute four auxiliary quantities k_1, k_2, k_3, k_4 and then the new value y_{n+1}. The method is well suited to the computer because it needs no special starting procedure, makes light demand on storage, and repeatedly uses the same straightforward computational procedure. It is numerically stable.

Note that, if f depends only on x, this method reduces to Simpson's rule of integration (Sec. 19.5). Note further that $k_1, \cdots, k_4$ depend on n and generally change from step to step.

Table 21.3 Classical Runge–Kutta Method of Fourth Order

ALGORITHM RUNGE–KUTTA (f, x_0, y_0, h, N).

This algorithm computes the solution of the initial value problem $y' = f(x, y), y(x_0) = y_0$ at equidistant points

(9) $$x_1 = x_0 + h, x_2 = x_0 + 2h, \cdots, x_N = x_0 + Nh;$$

here f is such that this problem has a unique solution on the interval $[x_0, x_N]$ (see Sec. 1.7).

INPUT: Function f, initial values x_0, y_0, step size h, number of steps N

OUTPUT: Approximation y_{n+1} to the solution $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$, where $n = 0, 1, \cdots, N - 1$

For $n = 0, 1, \cdots, N - 1$ do:

$k_1 = hf(x_n, y_n)$

$k_2 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_1)$

$k_3 = hf(x_n + \frac{1}{2}h, y_n + \frac{1}{2}k_2)$

$k_4 = hf(x_n + h, y_n + k_3)$

$x_{n+1} = x_n + h$

$y_{n+1} = y_n + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4)$

OUTPUT x_{n+1}, y_{n+1}

End

Stop

End RUNGE–KUTTA

[1]Named after the German mathematicians KARL RUNGE (Sec. 19.4) and WILHELM KUTTA (1867–1944). Runge [*Math. Annalen* **46** (1895), 167–178], the German mathematician KARL HEUN (1859–1929) [*Zeitschr. Math. Phys.* **45** (1900), 23–38], and Kutta [*Zeitschr. Math. Phys.* **46** (1901), 435–453] developed various similar methods. Theoretically, there are infinitely many fourth-order methods using four function values per step. The method in Table 21.3 is most popular from a practical viewpoint because of its "symmetrical" form and its simple coefficients. It was given by Kutta.

EXAMPLE 2 **Classical Runge–Kutta Method**

Apply the Runge–Kutta method to the initial value problem in Example 1, choosing $h = 0.2$, as before, and computing five steps.

Solution. For the present problem we have $f(x, y) = x + y$. Hence

$$k_1 = 0.2(x_n + y_n), \qquad k_2 = 0.2(x_n + 0.1 + y_n + 0.5k_1),$$
$$k_3 = 0.2(x_n + 0.1 + y_n + 0.5k_2), \qquad k_4 = 0.2(x_n + 0.2 + y_n + k_3).$$

Table 21.4 shows the results and their errors, which are smaller by factors 10^3 and 10^4 than those for the two Euler methods. See also Table 21.5. We mention in passing that since the present $k_1, \cdots, k_4$ are simple, operations were saved by substituting k_1 into k_2, then k_2 into k_3, etc.; the resulting formula is shown in Column 4 of Table 21.4. Keep in mind that we have four function evaluations at each step. ■

Table 21.4 Runge–Kutta Method Applied to (4)

n	x_n	y_n	$0.2214(x_n + y_n) + 0.0214$	Exact Values (6D) $y = e^x - x - 1$	$10^6 \times$ Error of y_n
0	0.0	0	0.021400	0.000000	0
1	0.2	0.021400	0.070418	0.021403	3
2	0.4	0.091818	0.130289	0.091825	7
3	0.6	0.222107	0.203414	0.222119	12
4	0.8	0.425521	0.292730	0.425541	20
5	1.0	0.718251		0.718282	31

Table 21.5 Comparison of the Accuracy of the Three Methods under Consideration in the Case of the Initial Value Problem (4), with $h = 0.2$

x	$y = e^x - x - 1$	Error		
		Euler (Table 21.1)	Improved Euler (Table 21.3)	Runge–Kutta (Table 21.5)
0.2	0.021403	0.021	0.0014	0.000003
0.4	0.091825	0.052	0.0034	0.000007
0.6	0.222119	0.094	0.0063	0.000011
0.8	0.425541	0.152	0.0102	0.000020
1.0	0.718282	0.230	0.0156	0.000031

Error and Step Size Control. RKF (Runge–Kutta–Fehlberg)

The idea of adaptive integration (Sec. 19.5) has analogs for Runge–Kutta (and other) methods. In Table 21.3 for RK (Runge–Kutta), if we compute in each step approximations $\tilde{y}$ and $\tilde{\tilde{y}}$ with step sizes h and $2h$, respectively, the latter has error per step equal to $2^5 = 32$ times that of the former; however, since we have only half as many steps for $2h$, the actual factor is $2^5/2 = 16$, so that, say,

$$\epsilon^{(2h)} \approx 16\epsilon^{(h)} \qquad \text{and thus} \qquad y^{(h)} - y^{(2h)} = \epsilon^{(2h)} - \epsilon^{(h)} \approx (16 - 1)\epsilon^{(h)}.$$

Hence the error $\epsilon = \epsilon^{(h)}$ for step size h is about

$$\epsilon = \tfrac{1}{15}(\tilde{y} - \tilde{\tilde{y}}) \tag{10}$$

where $\tilde{y} - \tilde{\tilde{y}} = y^{(h)} - y^{(2h)}$, as said before. Table 21.6 illustrates (10) for the initial value problem

$$y' = (y - x - 1)^2 + 2, \qquad y(0) = 1, \tag{11}$$

the step size $h = 0.1$ and $0 \leqq x \leqq 0.4$. We see that the estimate is close to the actual error. This method of error estimation is simple but may be unstable.

Table 21.6 Runge–Kutta Method Applied to the Initial Value Problem (11) and Error Estimate (10). Exact Solution $y = \tan x + x + 1$

x	$\tilde{y}$ (Step size h)	$\tilde{\tilde{y}}$ (Step size $2h$)	Error Estimate (10)	Actual Error	Exact Solution (9D)
0.0	1.000000000	1.000000000	0.000000000	0.000000000	1.000000000
0.1	1.200334589			0.000000083	1.200334672
0.2	1.402709878	1.402707408	0.000000165	0.000000157	1.402710036
0.3	1.609336039			0.000000210	1.609336250
0.4	1.822792993	1.822788993	0.000000267	0.000000226	1.822793219

RKF. E. Fehlberg [*Computing* **6** (1970), 61–71] proposed and developed error control by using two RK methods of different orders to go from (x_n, y_n) to (x_{n+1}, y_{n+1}). The difference of the computed y-values at x_{n+1} gives an error estimate to be used for step size control. Fehlberg discovered two RK formulas that together need only six function evaluations per step. We present these formulas here because RKF has become quite popular. For instance, Maple uses it (also for systems of ODEs).

Fehlberg's fifth-order RK method is

$$y_{n+1} = y_n + \gamma_1 k_1 + \cdots + \gamma_6 k_6 \tag{12a}$$

with coefficient vector $\gamma = [\gamma_1 \cdots \gamma_6]$,

$$\gamma = [\tfrac{16}{135} \quad 0 \quad \tfrac{6656}{12,825} \quad \tfrac{28,561}{56,430} \quad -\tfrac{9}{50} \quad \tfrac{2}{55}]. \tag{12b}$$

His **fourth-order RK method** is

$$y^*_{n+1} = y_n + \gamma^*_1 k_1 + \cdots + \gamma^*_5 k_5 \tag{13a}$$

with coefficient vector

$$\gamma^* = [\tfrac{25}{216} \quad 0 \quad \tfrac{1408}{2565} \quad \tfrac{2197}{4104} \quad -\tfrac{1}{5}]. \tag{13b}$$

In both formulas we use only six different function evaluations altogether, namely,

$$
\begin{aligned}
k_1 &= hf(x_n, y_n) \\
k_2 &= hf(x_n + \tfrac{1}{4}h, \quad y_n + \tfrac{1}{4}k_1) \\
k_3 &= hf(x_n + \tfrac{3}{8}h, \quad y_n + \tfrac{3}{32}k_1 + \tfrac{9}{32}k_2) \\
k_4 &= hf(x_n + \tfrac{12}{13}h, \quad y_n + \tfrac{1932}{2197}k_1 - \tfrac{7200}{2197}k_2 + \tfrac{7296}{2197}k_3) \\
k_5 &= hf(x_n + h, \quad y_n + \tfrac{439}{216}k_1 - 8k_2 + \tfrac{3680}{513}k_3 - \tfrac{845}{4104}k_4) \\
k_6 &= hf(x_n + \tfrac{1}{2}h, \quad y_n - \tfrac{8}{27}k_1 + 2k_2 - \tfrac{3544}{2565}k_3 + \tfrac{1859}{4104}k_4 - \tfrac{11}{40}k_5).
\end{aligned} \tag{14}
$$

The difference of (12) and (13) gives the **error estimate**

$$
\epsilon_{n+1} \approx y_{n+1} - y^*_{n+1} = \tfrac{1}{360}k_1 - \tfrac{128}{4275}k_3 - \tfrac{2197}{75{,}240}k_4 + \tfrac{1}{50}k_5 + \tfrac{2}{55}k_6. \tag{15}
$$

EXAMPLE 3 Runge–Kutta–Fehlberg

For the initial value problem (11) we obtain from (12)–(14) with $h = 0.1$ in the first step the 12S-values

$$
\begin{aligned}
k_1 &= 0.200000000000 & k_2 &= 0.200062500000 \\
k_3 &= 0.200140756867 & k_4 &= 0.200856926154 \\
k_5 &= 0.201006676700 & k_6 &= 0.200250418651
\end{aligned}
$$

$$
\begin{aligned}
y_1^* &= 1.20033466949 \\
y_1 &= 1.20033467253
\end{aligned}
$$

and the error estimate

$$
\epsilon_1 \approx y_1 - y_1^* = 0.00000000304.
$$

The exact 12S-value is $y(0.1) = 1.20033467209$. Hence the actual error of y_1 is $-4.4 \cdot 10^{-10}$, smaller than that in Table 21.6 by a factor of 200. ■

Table 21.7 summarizes essential features of the methods in this section. It can be shown that these methods are *numerically stable* (definition in Sec. 19.1). They are **one-step methods** because in each step we use the data of just *one* preceding step, in contrast to **multistep methods** where in each step we use data from *several* preceding steps, as we shall see in the next section.

Table 21.7 Methods Considered and Their Order (= Their Global Error)

Method	Function Evaluation per Step	Global Error	Local Error
Euler	1	$O(h)$	$O(h^2)$
Improved Euler	2	$O(h^2)$	$O(h^3)$
RK (fourth order)	4	$O(h^4)$	$O(h^5)$
RKF	6	$O(h^5)$	$O(h^6)$

Backward Euler Method. Stiff ODEs

The **backward Euler formula** for numerically solving (1) is

(16) $$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \qquad (n = 0, 1, \cdots).$$

This formula is obtained by evaluating the right side at the *new* location (x_{n+1}, y_{n+1}); this is called the **backward Euler scheme**. For known y_n it gives y_{n+1} *implicitly,* so it defines an **implicit method**, in contrast to the Euler method (3), which gives y_{n+1} explicitly. Hence (16) must be solved for y_{n+1}. How difficult this is depends on f in (1). For a linear ODE this provides no problem, as Example 4 (below) illustrates. The method is particularly useful for "stiff" ODEs, as they occur quite frequently in the study of vibrations, electric circuits, chemical reactions, etc. The situation of stiffness is roughly as follows; for details, see, for example, [E5], [E25], [E26] in App. 1.

Error terms of the methods considered so far involve a higher derivative. And we ask what happens if we let *h increase.* Now if the error (the derivative) grows fast but the desired solution also grows fast, nothing will happen. However, if that solution does not grow fast, then with growing h the error term can take over to an extent that the numeric result becomes completely nonsensical, as in Fig. 451. Such an ODE for which h must thus be restricted to small values, and the physical system the ODE models, are called **stiff**. This term is suggested by a mass–spring system with a stiff spring (spring with a large k; see Sec. 2.4). Example 4 illustrates that implicit methods remove the difficulty of increasing h in the case of stiffness: It can be shown that in the application of an implicit method the solution remains stable under any increase of h, although the accuracy decreases with increasing h.

EXAMPLE 4 Backward Euler Method. Stiff ODE

The initial value problem

$$y' = f(x, y) = -20hy + 20x^2 + 2x, \quad y(0) = 1$$

has the solution (verify!)

$$y = e^{-20x} + x^2.$$

The backward Euler formula (16) is

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) = y_n + h(-20y_{n+1} + 20x_{n+1}^2 + 2x_{n+1}).$$

Noting that $x_{n+1} = x_n + h$, taking the term $-20y_{n+1}$ to the left, and dividing, we obtain

(16*) $$y_{n+1} = \frac{y_n + h[20(x_n + h)^2 + 2(x_n + h)]}{1 + 20h}.$$

The numeric results in Table 21.8 show the following.

Stability of the backward Euler method for $h = 0.05$ and also for $h = 0.2$ with an error increase by about a factor 4 for $h = 0.2$,

Stability of the Euler method for $h = 0.05$ but instability for $h = 0.1$ (Fig. 451),

Stability of RK for $h = 0.1$ but instability for $h = 0.2$.

This illustrates that the ODE is stiff. Note that even in the case of stability the approximation of the solution near $x = 0$ is poor. ■

Stiffness will be considered further in Sec. 21.3 in connection with systems of ODEs.

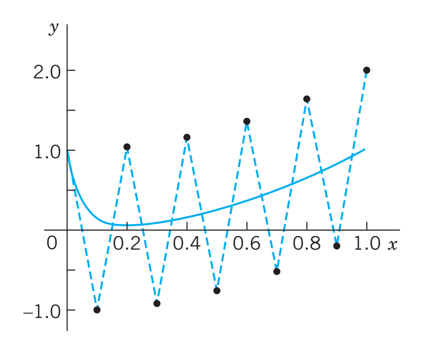


Fig. 451. Euler method with $h = 0.1$ for the stiff ODE in Example 4 and exact solution

Table 21.8 Backward Euler Method (BEM) for Example 6. Comparison with Euler and RK

x	BEM $h = 0.05$	BEM $h = 0.2$	Euler $h = 0.05$	Euler $h = 0.1$	RK $h = 0.1$	RK $h = 0.2$	Exact
0.0	1.00000	1.00000	1.00000	1.00000	1.00000	1.000	1.00000
0.1	0.26188		0.00750	−1.00000	0.34500		0.14534
0.2	0.10484	0.24800	0.03750	1.04000	0.15333	5.093	0.05832
0.3	0.10809		0.08750	−0.92000	0.12944		0.09248
0.4	0.16640	0.20960	0.15750	1.16000	0.17482	25.48	0.16034
0.5	0.25347		0.24750	−0.76000	0.25660		0.25004
0.6	0.36274	0.37792	0.35750	1.36000	0.36387	127.0	0.36001
0.7	0.49256		0.48750	−0.52000	0.49296		0.49001
0.8	0.64252	0.65158	0.63750	1.64000	0.64265	634.0	0.64000
0.9	0.81250		0.80750	−0.20000	0.81255		0.81000
1.0	1.00250	1.01032	0.99750	2.00000	1.00252	3168	1.00000

PROBLEM SET 21.1

1–4 EULER METHOD

Do 10 steps. Solve exactly. Compute the error. Show details.

1. $y' + 0.2y = 0, \quad y(0) = 5, \quad h = 0.2$
2. $y' = \frac{1}{2}\pi\sqrt{1 - y^2}, \quad y(0) = 0, \quad h = 0.1$
3. $y' = (y - x)^2, \quad y(0) = 0, \quad h = 0.1$
4. $y' = (y + x)^2, \quad y(0) = 0, \quad h = 0.1$

5–10 IMPROVED EULER METHOD

Do 10 steps. Solve exactly. Compute the error. Show details.

5. $y' = y, \quad y(0) = 1, \quad h = 0.1$
6. $y' = 2(1 + y^2), \quad y(0) = 0, \quad h = 0.05$
7. $y' - xy^2 = 0, \quad y(0) = 1, \quad h = 0.1$
8. **Logistic population model.** $y' = y - y^2, \quad y(0) = 0.2, \quad h = 0.1$
9. Do Prob. 7 using Euler's method with $h = 0.1$ and compare the accuracy.
10. Do Prob. 7 using the improved Euler method, 20 steps with $h = 0.05$. Compare.

11–17 CLASSICAL RUNGE–KUTTA METHOD OF FOURTH ORDER

Do 10 steps. Compare as indicated. Show details.

11. $y' - xy^2 = 0, \quad y(0) = 1, \quad h = 0.1$. Compare with Prob. 7. Apply the error estimate (10) to y_{10}.
12. $y' = y - y^2, \quad y(0) = 0.2, \quad h = 0.1$. Compare with Prob. 8.
13. $y' = 1 + y^2, \quad y(0) = 0, \quad h = 0.1$
14. $y' = (1 - x^{-1})y, \quad y(1) = 1, \quad h = 0.1$
15. $y' + y \tan x = \sin 2x, \quad y(0) = 1, \quad h = 0.1$
16. Do Prob. 15 with $h = 0.2$, 5 steps, and compare the errors with those in Prob. 15.

17. $y' = 4x^3y^2, \quad y(0) = 0.5, \quad h = 0.1$

18. Kutta's third-order method is defined by $y_{n+1} = y_n + \frac{1}{6}(k_1 + 4k_2 + k_3^*)$ with k_1 and k_2 as in RK (Table 21.3) and $k_3^* = hf(x_{n+1}, y_n - k_1 + 2k_2)$. Apply this method to (4) in (6). Choose $h = 0.2$ and do 5 steps. Compare with Table 21.5.

19. CAS EXPERIMENT. Euler–Cauchy vs. RK. Consider the initial value problem

$$(17) \quad y' = (y - 0.01x^2)^2 \sin(x^2) + 0.02x, \qquad y(0) = 0.4$$

(solution: $y = 1/[2.5 - S(x)] + 0.01x^2$ where $S(x)$ is the Fresnel integral (38) in App. 3.1).

(a) Solve (17) by Euler, improved Euler, and RK methods for $0 \leqq x \leqq 5$ with step $h = 0.2$. Compare the errors for $x = 1, 3, 5$ and comment.

(b) Graph solution curves of the ODE in (17) for various positive and negative initial values.

(c) Do a similar experiment as in (a) for an initial value problem that has a monotone increasing or monotone decreasing solution. Compare the behavior of the error with that in (a). Comment.

20. CAS EXPERIMENT. RKF. (a) Write a program for RKF that gives x_n, y_n, the estimate (10), and, if the solution is known, the actual error ϵ_n.

(b) Apply the program to Example 3 in the text (10 steps, $h = 0.1$).

(c) ϵ_n in (b) gives a relatively good idea of the size of the actual error. Is this typical or accidental? Find out, by experimentation with other problems, on what properties of the ODE or solution this might depend.

21.2 Multistep Methods

In a **one-step method** we compute y_{n+1} using only a single step, namely, the previous value y_n. *One-step methods are* "**self-starting**," they need no help to get going because they obtain y_1 from the initial value y_0, etc. All methods in Sec. 21.1 are one-step.

In contrast, a **multistep method** uses, in each step, values from two or more previous steps. These methods are motivated by the expectation that the additional information will increase accuracy and stability. But to get started, one needs values, say, y_0, y_1, y_2, y_3 in a 4-step method, obtained by Runge–Kutta or another accurate method. Thus, multistep methods are not self-starting. Such methods are obtained as follows.

Adams–Bashforth Methods

We consider an initial value problem

$$\textbf{(1)} \qquad y' = f(x, y), \qquad y(x_0) = y_0$$

as before, with f such that the problem has a unique solution on some open interval containing x_0. We integrate $y' = f(x, y)$ from x_n to $x_{n+1} = x_n + h$. This gives

$$\int_{x_n}^{x_{n+1}} y'(x)\,dx = y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} f(x, y(x))\,dx.$$

Now comes the main idea. We replace $f(x, y(x))$ by an interpolation polynomial $p(x)$ (see Sec. 19.3), so that we can later integrate. This gives approximations y_{n+1} of $y(x_{n+1})$ and y_n of $y(x_n)$,

$$(2) \qquad y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} p(x)\,dx.$$

Different choices of $p(x)$ will now produce different methods. We explain the principle by taking a cubic polynomial, namely, the polynomial $p_3(x)$ that at (equidistant)

$$x_n, \qquad x_{n-1}, \qquad x_{n-2}, \qquad x_{n-3}$$

has the respective values

$$(3) \qquad \begin{aligned} f_n &= f(x_n, y_n) \\ f_{n-1} &= f(x_{n-1}, y_{n-1}) \\ f_{n-2} &= f(x_{n-2}, y_{n-2}) \\ f_{n-3} &= f(x_{n-3}, y_{n-3}). \end{aligned}$$

This will lead to a practically useful formula. We can obtain $p_3(x)$ from Newton's backward difference formula (18), Sec. 19.3:

$$p_3(x) = f_n + r\nabla f_n + \tfrac{1}{2}r(r+1)\nabla^2 f_n + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_n$$

where

$$r = \frac{x - x_n}{h}.$$

We integrate $p_3(x)$ over x from x_n to $x_{n+1} = x_n + h$, thus over r from 0 to 1. Since

$$x = x_n + hr, \qquad \text{we have} \qquad dx = h\,dr.$$

The integral of $\frac{1}{2}r(r+1)$ is $\frac{5}{12}$ and that of $\frac{1}{6}r(r+1)(r+2)$ is $\frac{3}{8}$. We thus obtain

$$(4) \qquad \int_{x_n}^{x_{n+1}} p_3\,dx = h\int_0^1 p_3\,dr = h\left(f_n + \frac{1}{2}\nabla f_n + \frac{5}{12}\nabla^2 f_n + \frac{3}{8}\nabla^3 f_n\right).$$

It is practical to replace these differences by their expressions in terms of f:

$$\begin{aligned} \nabla f_n &= f_n - f_{n-1} \\ \nabla^2 f_n &= f_n - 2f_{n-1} + f_{n-2} \\ \nabla^3 f_n &= f_n - 3f_{n-1} + 3f_{n-2} - f_{n-3}. \end{aligned}$$

We substitute this into (4) and collect terms. This gives the multistep formula of the **Adams–Bashforth method**[2] *of fourth order*

$$\textbf{(5)} \qquad y_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}).$$

[2]Named after JOHN COUCH ADAMS (1819–1892), English astronomer and mathematician, one of the predictors of the existence of the planet Neptune (using mathematical calculations), director of the Cambridge Observatory; and FRANCIS BASHFORTH (1819–1912), English mathematician.

It expresses the new value y_{n+1} [approximation of the solution y of (1) at x_{n+1}] in terms of 4 values of f computed from the y-values obtained in the preceding 4 steps. The local truncation error is of order h^5, as can be shown, so that the global error is of order h^4; hence (5) does define a fourth-order method.

Adams–Moulton Methods

Adams–Moulton methods are obtained if for $p(x)$ in (2) we choose a polynomial that interpolates $f(x, y(x))$ at $x_{n+1}, x_n, x_{n-1}, \cdots$ (as opposed to $x_n, x_{n-1}, \cdots$ used before; this is the main point). We explain the principle for the cubic polynomial $\tilde{p}_3(x)$ that interpolates at $x_{n+1}, x_n, x_{n-1}, x_{n-2}$. (Before we had $x_n, x_{n-1}, x_{n-2}, x_{n-3}$.) Again using (18) in Sec. 19.3 but now setting $r = (x - x_{n+1})/h$, we have

$$\tilde{p}_3(x) = f_{n+1} + r\nabla f_{n+1} + \tfrac{1}{2}r(r+1)\nabla^2 f_{n+1} + \tfrac{1}{6}r(r+1)(r+2)\nabla^3 f_{n+1}.$$

We now integrate over x from x_n to x_{n+1} as before. This corresponds to integrating over r from -1 to 0. We obtain

$$\int_{x_n}^{x_{n+1}} \tilde{p}_3(x)\,dx = h\left(f_{n+1} - \frac{1}{2}\nabla f_{n+1} - \frac{1}{12}\nabla^2 f_{n+1} - \frac{1}{24}\nabla^3 f_{n+1}\right).$$

Replacing the differences as before gives

$$\text{(6)} \qquad y_{n+1} = y_n + \int_{x_n}^{x_{n+1}} \tilde{p}_3(x)\,dx = y_n + \frac{h}{24}(9f_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

This is usually called an **Adams–Moulton formula.**[3] It is an **implicit formula** because $f_{n+1} = f(x_{n+1}, y_{n+1})$ appears on the right, so that it defines y_{n+1} only *implicitly,* in contrast to (5), which is an **explicit formula**, not involving y_{n+1} on the right. To use (6) we must ***predict*** a value y^*_{n+1}, for instance, by using (5), that is,

$$\textbf{(7a)} \qquad y^*_{n+1} = y_n + \frac{h}{24}(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}).$$

The ***corrected*** new value y_{n+1} is then obtained from (6) with f_{n+1} replaced by $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$ and the other f's as in (6); thus,

$$\textbf{(7b)} \qquad y_{n+1} = y_n + \frac{h}{24}(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}).$$

This **predictor–corrector method** (7a), (7b) is usually called the **Adams–Moulton method** *of fourth order.* It has the advantage over RK that (7) gives the error estimate

$$\epsilon_{n+1} \approx \tfrac{1}{15}(y_{n+1} - y^*_{n+1}),$$

as can be shown. This is the analog of (10) in Sec. 21.1.

[3] FOREST RAY MOULTON (1872–1952), American astronomer at the University of Chicago. For ADAMS see footnote 2.

Sometimes the name Adams–Moulton method is reserved for the method with *several* corrections per step by (7b) until a specific accuracy is reached. Popular codes exist for both versions of the method.

Getting Started. In (5) we need f_0, f_1, f_2, f_3. Hence from (3) we see that we must first compute y_1, y_2, y_3 by some other method of comparable accuracy, for instance, by RK or by RKF. For other choices see Ref. [E26] listed in App. 1.

EXAMPLE 1 Adams–Bashforth Prediction (7a), Adams–Moulton Correction (7b)

Solve the initial value problem

$$y' = x + y, \qquad y(0) = 0 \tag{8}$$

by (7a), (7b) on the interval $0 \leqq x \leqq 2$, choosing $h = 0.2$.

Solution. The problem is the same as in Examples 1 and 2, Sec. 21.1, so that we can compare the results. We compute starting values y_1, y_2, y_3 by the classical Runge–Kutta method. Then in each step we predict by (7a) and make one correction by (7b) before we execute the next step. The results are shown and compared with the exact values in Table 21.9. We see that the corrections improve the accuracy considerably. This is typical. ■

Table 21.9 Adams–Moulton Method Applied to the Initial Value Problem (8); Predicted Values Computed by (7a) and Corrected Values by (7b)

n	x_n	Starting y_n	Predicted y_n^*	Corrected y_n	Exact Values	$10^6 \cdot$ Error of y_n
0	0.0	0.000000			0.000000	0
1	0.2	0.021400			0.021403	3
2	0.4	0.091818			0.091825	7
3	0.6	0.222107			0.222119	12
4	0.8		0.425361	0.425529	0.425541	12
5	1.0		0.718066	0.718270	0.718282	12
6	1.2		1.119855	1.120106	1.120117	11
7	1.4		1.654885	1.655191	1.655200	9
8	1.6		2.352653	2.353026	2.353032	6
9	1.8		3.249190	3.249646	3.249647	1
10	2.0		4.388505	4.389062	4.389056	−6

Comments on Comparison of Methods. An Adams–Moulton formula is generally much more accurate than an Adams–Bashforth formula of the same order. This justifies the greater complication and expense in using the former. The method (7a), (7b) is ***numerically stable***, whereas the exclusive use of (7a) might cause instability. Step size control is relatively simple. If $|\text{Corrector} - \text{Predictor}| > \text{TOL}$, use interpolation to generate "old" results at half the current step size and then try $h/2$ as the new step.

Whereas the Adams–Moulton formula (7a), (7b) needs only 2 evaluations per step, Runge–Kutta needs 4; however, with Runge–Kutta one may be able to take a step size more than twice as large, so that a comparison of this kind (widespread in the literature) is meaningless.

For more details, see Refs. [E25], [E26] listed in App. 1.

PROBLEM SET 21.2

1–10 ADAMS–MOULTON METHOD

Solve the initial value problem by Adams–Moulton (7a), (7b), 10 steps with 1 correction per step. Solve exactly and compute the error. Use RK where no starting values are given.

1. $y' = y$, $y(0) = 1$, $h = 0.1$, (1.105171, 1.221403, 1.349858)
2. $y' = 2xy$, $y(0) = 1$, $h = 0.1$
3. $y' = 1 + y^2$, $y(0) = 0$, $h = 0.1$, (0.100335, 0.202710, 0.309336)
4. Do Prob. 2 by RK, 5 steps, $h = 0.2$. Compare the errors.
5. Do Prob. 3 by RK, 5 steps, $h = 0.2$. Compare the errors.
6. $y' = (y - x - 1)^2 + 2$, $y(0) = 1$, $h = 0.1$, 10 steps
7. $y' = 3y - 12y^2$, $y(0) = 0.2$, $h = 0.1$
8. $y' = 1 - 4y^2$, $y(0) = 0$, $h = 0.1$
9. $y' = 3x^2(1 + y)$, $y(0) = 0$, $h = 0.05$
10. $y' = x/y$, $y(1) = 3$, $h = 0.2$
11. Do and show the calculations leading to (4)–(7) in the text.
12. **Quadratic polynomial.** Apply the method in the text to a polynomial of second degree. Show that this leads to the predictor and corrector formulas

$$y^*_{n+1} = y_n + \frac{h}{12}(23f_n - 16f_{n-1} + 5f_{n-2}),$$
$$y_{n+1} = y_n + \frac{h}{12}(5f_{n+1} + 8f_n - f_{n-1}).$$

13. Using Prob. 12, solve $y' = 2xy$, $y(0) = 1$ (10 steps, $h = 0.1$, RK starting values). Compare with the exact solution and comment.
14. How much can you reduce the error in Prob. 13 by halfing h (20 steps, $h = 0.05$)? First guess, then compute.
15. **CAS PROJECT. Adams–Moulton. (a) Accurate starting** is important in (7a), (7b). Illustrate this in Example 1 of the text by using starting values from the improved Euler–Cauchy method and compare the results with those in Table 21.8.

(b) How much does the error in Prob. 11 decrease if you use exact starting values (instead of RK values)?

(c) Experiment to find out for what ODEs poor starting is very damaging and for what ODEs it is not.

(d) The classical **RK method** often gives the same accuracy with step $2h$ as Adams–Moulton with step h, so that the total number of function evaluations is the same in both cases. Illustrate this with Prob. 8. (Hence corresponding comparisons in the literature in favor of Adams–Moulton are not valid. See also Probs. 6 and 7.)

21.3 Methods for Systems and Higher Order ODEs

Initial value problems for first-order systems of ODEs are of the form

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \mathbf{y}(x_0) = \mathbf{y}_0, \tag{1}$$

in components

$$\begin{aligned} y_1' &= f_1(x, y_1, \cdots, y_m), & y_1(x_0) &= y_{10} \\ y_2' &= f_2(x, y_1, \cdots, y_m), & y_2(x_0) &= y_{20} \\ &\cdots\cdots\cdots\cdots\cdots\cdots & &\cdots\cdots\cdots\cdots \\ y_m' &= f_m(x, y_1, \cdots, y_m). & y_m(x_0) &= y_{m0}. \end{aligned}$$

Here, $\mathbf{f}$ is assumed to be such that the problem has a unique solution $\mathbf{y}(x)$ on some open x-interval containing x_0. Our discussion will be independent of Chap. 4 on systems.

Before explaining solution methods it is important to note that (1) includes initial value problems for single mth-order ODEs,

$$y^{(m)} = f(x, y, y', y'', \cdots, y^{(m-1)}) \tag{2}$$

and initial conditions $y(x_0) = K_1, y'(x_0) = K_2, \cdots, y^{(m-1)}(x_0) = K_m$ as special cases.

Indeed, the connection is achieved by setting

$$y_1 = y, \qquad y_2 = y', \qquad y_3 = y'', \qquad \cdots, \qquad y_m = y^{(m-1)}. \tag{3}$$

Then we obtain the system

$$\begin{aligned} y_1' &= y_2 \\ y_2' &= y_3 \\ &\vdots \\ y_{m-1}' &= y_m \\ y_m' &= f(x, y_1, \cdots, y_m) \end{aligned} \tag{4}$$

and the initial conditions $y_1(x_0) = K_1, \quad y_2(x_0) = K_2, \quad \cdots, \quad y_m(x_0) = K_m$.

Euler Method for Systems

Methods for single first-order ODEs can be extended to systems (1) simply by writing vector functions $\mathbf{y}$ and $\mathbf{f}$ instead of scalar functions y and f, whereas x remains a scalar variable.

We begin with the Euler method. Just as for a single ODE, this method will not be accurate enough for practical purposes, but it nicely illustrates the extension principle.

EXAMPLE 1 Euler Method for a Second-Order ODE. Mass–Spring System

Solve the initial value problem for a damped mass–spring system

$$y'' + 2y' + 0.75y = 0, \qquad y(0) = 3, \qquad y'(0) = -2.5$$

by the Euler method for systems with step $h = 0.2$ for x from 0 to 1 (where x is time).

Solution. The **Euler method** (3), Sec. 21.1, generalizes to systems in the form

$$\mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_n, \mathbf{y}_n), \tag{5}$$

in components

$$\begin{aligned} y_{1,n+1} &= y_{1,n} + hf_1(x_n, y_{1,n}, y_{2,n}) \\ y_{2,n+1} &= y_{2,n} + hf_2(x_n, y_{1,n}, y_{2,n}) \end{aligned}$$

and similarly for systems of more than two equations. By (4) the given ODE converts to the system

$$\begin{aligned} y_1' &= f_1(x, y_1, y_2) = y_2 \\ y_2' &= f_2(x, y_1, y_2) = -2y_2 - 0.75y_1. \end{aligned}$$

Hence (5) becomes

$$y_{1,n+1} = y_{1,n} + 0.2y_{2,n}$$
$$y_{2,n+1} = y_{2,n} + 0.2(-2y_{2,n} - 0.75y_{1,n}).$$

The initial conditions are $y(0) = y_1(0) = 3$, $y'(0) = y_2(0) = -2.5$. The calculations are shown in Table 21.10. As for single ODEs, the results would not be accurate enough for practical purposes. The example merely serves to illustrate the method because the problem can be readily solved exactly,

$$y = y_1 = 2e^{-0.5x} + e^{-1.5x}, \qquad \text{thus} \qquad y' = y_2 = -e^{-0.5x} - 1.5e^{-1.5x}.$$ ■

Table 21.10 Euler Method for Systems in Example 1 (Mass–Spring System)

n	x_n	$y_{1,n}$	y_1 Exact (5D)	Error $\epsilon_1 = y_1 - y_{1,n}$	$y_{2,n}$	y_2 Exact (5D)	Error $\epsilon_2 = y_2 - y_{2,n}$
0	0.0	3.00000	3.00000	0.00000	−2.50000	−2.50000	0.00000
1	0.2	2.50000	2.55049	0.05049	−1.95000	−2.01606	−0.06606
2	0.4	2.11000	2.18627	0.76270	−1.54500	−1.64195	−0.09695
3	0.6	1.80100	1.88821	0.08721	−1.24350	−1.35067	−0.10717
4	0.8	1.55230	1.64183	0.08953	−1.01625	−1.12211	−0.10586
5	1.0	1.34905	1.43619	0.08714	−0.84260	−0.94123	−0.09863

Runge–Kutta Methods for Systems

As for Euler methods, we obtain RK methods for an initial value problem (1) simply by writing vector formulas for vectors with m components, which, for $m = 1$, reduce to the previous scalar formulas.

Thus, for the *classical* **RK method** *of fourth order* in Table 21.3, we obtain

(6a) $$\mathbf{y}(x_0) = \mathbf{y}_0 \qquad \text{(Initial values)}$$

and for each step $n = 0, 1, \cdots, N - 1$ we obtain the 4 auxiliary quantities

(6b)
$$\begin{aligned}
\mathbf{k}_1 &= h\mathbf{f}(x_n, & &\mathbf{y}_n) \\
\mathbf{k}_2 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, & &\mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_1) \\
\mathbf{k}_3 &= h\mathbf{f}(x_n + \tfrac{1}{2}h, & &\mathbf{y}_n + \tfrac{1}{2}\mathbf{k}_2) \\
\mathbf{k}_4 &= h\mathbf{f}(x_n + h, & &\mathbf{y}_n + \mathbf{k}_3)
\end{aligned}$$

and the new value [approximation of the solution $\mathbf{y}(x)$ at $x_{n+1} = x_0 + (n + 1)h$]

(6c) $$\mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4).$$

EXAMPLE 2 RK Method for Systems. Airy's Equation. Airy Function Ai(x)

Solve the initial value problem

$$y'' = xy, \qquad y(0) = 1/(3^{2/3} \cdot \Gamma(\tfrac{2}{3})) = 0.35502805, \qquad y'(0) = -1/(3^{1/3} \cdot \Gamma(\tfrac{1}{3})) = -0.25881940$$

by the Runge–Kutta method for systems with $h = 0.2$; do 5 steps. This is **Airy's equation**,[4] which arose in optics (see Ref. [A13], p. 188, listed in App. 1). Γ is the gamma function (see App. A3.1). The initial conditions are such that we obtain a standard solution, the **Airy function** $\mathrm{Ai}(x)$, a special function that has been thoroughly investigated; for numeric values, see Ref. [GenRef1], pp. 446, 475.

Solution. For $y'' = xy$, setting $y_1 = y$, $y_2 = y_1' = y'$ we obtain the system (4)

$$y_1' = y_2$$
$$y_2' = xy_1.$$

Hence $\mathbf{f} = [f_1 \quad f_2]^{\mathsf{T}}$ in (1) has the components $f_1(x, y) = y_2$, $f_2(x, y) = xy_1$. We now write (6) in components. The initial conditions (6a) are $y_{1,0} = 0.35502805$, $y_{2,0} = -0.25881940$. In (6b) we have fewer subscripts by simply writing $\mathbf{k}_1 = \mathbf{a}$, $\mathbf{k}_2 = \mathbf{b}$, $\mathbf{k}_3 = \mathbf{c}$, $\mathbf{k}_4 = \mathbf{d}$, so that $\mathbf{a} = [a_1 \quad a_2]^{\mathsf{T}}$, etc. Then (6b) takes the form

$$\text{(6b*)} \qquad \begin{aligned} \mathbf{a} &= h\begin{bmatrix} y_{2,n} \\ x_n y_{1,n} \end{bmatrix} \\ \mathbf{b} &= h\begin{bmatrix} y_{2,n} + \tfrac{1}{2}a_2 \\ (x_n + \tfrac{1}{2}h)(y_{1,n} + \tfrac{1}{2}a_1) \end{bmatrix} \\ \mathbf{c} &= h\begin{bmatrix} y_{2,n} + \tfrac{1}{2}b_2 \\ (x_n + \tfrac{1}{2}h)(y_{1,n} + \tfrac{1}{2}b_1) \end{bmatrix} \\ \mathbf{d} &= h\begin{bmatrix} y_{2,n} + c_2 \\ (x_n + h)(y_{1,n} + c_1) \end{bmatrix}. \end{aligned}$$

For example, the second component of $\mathbf{b}$ is obtained as follows. $\mathbf{f}(x, y)$ has the second component $f_2(x, y) = xy_1$. Now in $\mathbf{b}$ $(= \mathbf{k}_2)$ the first argument is

$$x = x_n + \tfrac{1}{2}h.$$

The second argument in $\mathbf{b}$ is

$$\mathbf{y} = \mathbf{y}_n + \tfrac{1}{2}\mathbf{a},$$

and the first component of this is

$$y_1 = y_{1,n} + \tfrac{1}{2}a_1.$$

Together,

$$xy_1 = (x_n + \tfrac{1}{2}h)(y_{1,n} + \tfrac{1}{2}a_1).$$

Similarly for the other components in (6b*). Finally,

$$\text{(6c*)} \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + \tfrac{1}{6}(\mathbf{a} + 2\mathbf{b} + 2\mathbf{c} + \mathbf{d}).$$

Table 21.11 shows the values $y(x) = y_1(x)$ of the Airy function $\mathrm{Ai}(x)$ and of its derivative $y'(x) = y_2(x)$ as well as of the (rather small!) error of $y(x)$. ■

---

[4]Named after Sir GEORGE BIDELL AIRY (1801–1892), English mathematician, who is known for his work in elasticity and in PDEs.

Table 21.11 RK Method for Systems: Values $y_{1,n}(x_n)$ of the Airy Function Ai(x) in Example 2

n	x_n	$y_{1,n}(x_n)$	$y_1(x_n)$ Exact (8D)	$10^8 \cdot$ Error of y_1	$y_{2,n}(x_n)$
0	0.0	0.35502805	0.35502805	0	−0.25881940
1	0.2	0.30370303	0.30370315	12	−0.25240464
2	0.4	0.25474211	0.25474235	24	−0.23583073
3	0.6	0.20979973	0.20980006	33	−0.21279185
4	0.8	0.16984596	0.16984632	36	−0.18641171
5	1.0	0.13529207	0.13529242	35	−0.15914687

Runge–Kutta–Nyström Methods (RKN Methods)

RKN methods are direct extensions of RK methods (Runge–Kutta methods) to second-order ODEs $y'' = f(x, y, y')$, as given by the Finnish mathematician E. J. Nyström [*Acta Soc. Sci. fenn.,* 1925, L, No. 13]. The best known of these uses the following formulas, where $n = 0, 1, \cdots, N - 1$ (N the number of steps):

(7a)
$$\begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n, y'_n) \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y'_n + k_1) \qquad \text{where } K = \tfrac{1}{2}h(y'_n + \tfrac{1}{2}k_1) \\ k_3 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + K, y'_n + k_2) \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + L, y'_n + 2k_3) \qquad \text{where } L = h(y'_n + k_3). \end{aligned}$$

From this we compute the approximation y_{n+1} of $y(x_{n+1})$ at $x_{n+1} = x_0 + (n + 1)h$,

(7b)
$$y_{n+1} = y_n + h(y'_n + \tfrac{1}{3}(k_1 + k_2 + k_3)),$$

and the approximation y'_{n+1} of the derivative $y'(x_{n+1})$ needed in the next step,

(7c)
$$y'_{n+1} = y'_n + \tfrac{1}{3}(k_1 + 2k_2 + 2k_3 + k_4).$$

RKN for ODEs $y'' = f(x, y)$ Not Containing y'. Then $k_2 = k_3$ in (7), which makes the method particularly advantageous and reduces (7a)–(7c) to

(7*)
$$\begin{aligned} k_1 &= \tfrac{1}{2}hf(x_n, y_n) \\ k_2 &= \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}h(y'_n + \tfrac{1}{2}k_1)) = k_3 \\ k_4 &= \tfrac{1}{2}hf(x_n + h, y_n + h(y'_n + k_2)) \\ y_{n+1} &= y_n + h(y'_n + \tfrac{1}{3}(k_1 + 2k_2)) \\ y'_{n+1} &= y'_n + \tfrac{1}{3}(k_1 + 4k_2 + k_4). \end{aligned}$$

EXAMPLE 3 RKN Method. Airy's Equation. Airy Function Ai(x)

For the problem in Example 2 and $h = 0.2$ as before we obtain from (7*) simply $k_1 = 0.1x_n y_n$ and

$$k_2 = k_3 = 0.1(x_n + 0.1)(y_n + 0.1y'_n + 0.05k_1), \qquad k_4 = 0.1(x_n + 0.2)(y_n + 0.2y'_n + 0.2k_2).$$

Table 21.12 shows the results. The accuracy is the same as in Example 2, but the work was much less. ■

Table 21.12 Runge–Kutta–Nyström Method Applied to Airy's Equation, Computation of the Airy Function $y = \mathrm{Ai}(x)$

x_n	y_n	y_n'	$y(x)$ Exact (8D)	$10^8 \cdot$ Error of y_n
0.0	0.35502805	−0.25881940	0.35502805	0
0.2	0.30370304	−0.25240464	0.30370315	11
0.4	0.25474211	−0.23583070	0.25474235	24
0.6	0.20979974	−0.21279172	0.20980006	32
0.8	0.16984599	−0.18641134	0.16984632	33
1.0	0.13529218	−0.15914609	0.13529242	24

Our work in Examples 2 and 3 also illustrates that usefulness of methods for ODEs in the computation of values of "**higher transcendental functions**."

Backward Euler Method for Systems. Stiff Systems

The backward Euler formula (16) in Sec. 21.1 generalizes to systems in the form

$$\textbf{(8)} \qquad \mathbf{y}_{n+1} = \mathbf{y}_n + h\mathbf{f}(x_{n+1}, \mathbf{y}_{n+1}) \qquad (n = 0, 1, \cdots).$$

This is again an implicit method, giving $\mathbf{y}_{n+1}$ implicitly for given $\mathbf{y}_n$. Hence (8) must be solved for $\mathbf{y}_{n+1}$. For a linear system this is shown in the next example. This example also illustrates that, similar to the case of a single ODE in Sec. 21.1, the method is very useful for **stiff systems**. These are systems of ODEs whose matrix has eigenvalues λ of very different magnitudes, having the effect that, just as in Sec. 21.1, the step in direct methods, RK for example, cannot be increased beyond a certain threshold without losing stability. ($\lambda = -1$ and -10 in Example 4, but larger differences do occur in applications.)

EXAMPLE 4 Backward Euler Method for Systems of ODEs. Stiff Systems

Compare the backward Euler method (8) with the Euler and the RK methods for numerically solving the initial value problem

$$y'' + 11y' + 10y = 10x + 11, \qquad y(0) = 2, \qquad y'(0) = -10$$

converted to a system of first-order ODEs.

Solution. The given problem can easily be solved, obtaining

$$y = e^{-x} + e^{-10x} + x$$

so that we can compute errors. Conversion to a system by setting $y = y_1, y' = y_2$ [see (4)] gives

$$\begin{aligned} y_1' &= y_2 & y_1(0) &= 2 \\ y_2' &= -10y_1 - 11y_2 + 10x + 11 & y_2(0) &= -10. \end{aligned}$$

The coefficient matrix

$$\mathbf{A} = \begin{bmatrix} 0 & 1 \\ -10 & -11 \end{bmatrix} \quad \text{has the characteristic determinant} \quad \begin{vmatrix} -\lambda & 1 \\ -10 & -\lambda - 11 \end{vmatrix}$$

whose value is $\lambda^2 + 11\lambda + 10 = (\lambda + 1)(\lambda + 10)$. Hence the eigenvalues are -1 and -10 as claimed above. The backward Euler formula is

$$\mathbf{y}_{n+1} = \begin{bmatrix} y_{1,n+1} \\ y_{2,n+1} \end{bmatrix} = \begin{bmatrix} y_{1,n} \\ y_{2,n} \end{bmatrix} + h \begin{bmatrix} y_{2,n+1} \\ -10y_{1,n+1} - 11y_{2,n+1} + 10x_{n+1} + 11 \end{bmatrix}.$$

Reordering terms gives the linear system in the unknowns $y_{1,n+1}$ and $y_{2,n+1}$

$$\begin{aligned} y_{1,n+1} - \qquad hy_{2,n+1} &= y_{1,n} \\ 10hy_{1,n+1} + (1 + 11h)y_{2,n+1} &= y_{2,n} + 10h(x_n + h) + 11h. \end{aligned}$$

The coefficient determinant is $D = 1 + 11h + 10h^2$, and Cramer's rule (in Sec. 7.6) gives the solution

$$\mathbf{y}_{n+1} = \frac{1}{D}\begin{bmatrix} (1 + 11h)y_{1,n} + hy_{2,n} + 10h^2x_n + 11h^2 + 10h^3 \\ -10hy_{1,n} + \quad y_{2,n} + \quad 10hx_n + \quad 11h + 10h^2 \end{bmatrix}.$$

Table 21.13 Backward Euler Method (BEM) for Example 4. Comparison with Euler and RK

x	BEM $h = 0.2$	BEM $h = 0.4$	Euler $h = 0.1$	Euler $h = 0.2$	RK $h = 0.2$	RK $h = 0.3$	Exact
0.0	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000	2.00000
0.2	1.36667		1.01000	0.00000	1.35207		1.15407
0.4	1.20556	1.31429	1.56100	2.04000	1.18144		1.08864
0.6	1.21574		1.13144	0.11200	1.18585	3.03947	1.15129
0.8	1.29460	1.35020	1.23047	2.20960	1.26168		1.24966
1.0	1.40599		1.34868	0.32768	1.37200		1.36792
1.2	1.53627	1.57243	1.48243	2.46214	1.50257	5.07569	1.50120
1.4	1.67954		1.62877	0.60972	1.64706		1.64660
1.6	1.83272	1.86191	1.78530	2.76777	1.80205		1.80190
1.8	1.99386		1.95009	0.93422	1.96535	8.72329	1.96530
2.0	2.16152	2.18625	2.12158	3.10737	2.13536		2.13534

Table 21.13 shows the following.

Stability of the backward Euler method for $h = 0.2$ and 0.4 (and in fact for any h; try $h = 5.0$) with decreasing accuracy for increasing h

Stability of the Euler method for $h = 0.1$ but instability for $h = 0.2$

Stability of RK for $h = 0.2$ but instability for $h = 0.3$

Figure 452 shows the Euler method for $h = 0.18$, an interesting case with initial jumping (for about $x > 3$) but later monotone following the solution curve of $y = y_1$. See also CAS Experiment 15. ■

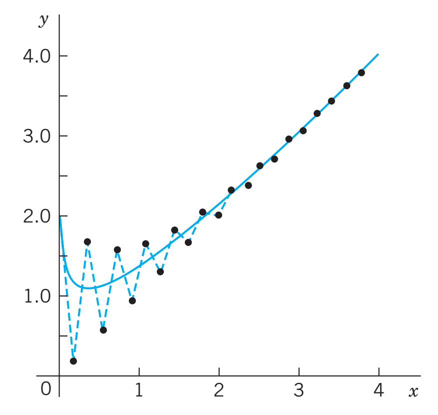


Fig. 452. Euler method with $h = 0.18$ in Example 4

PROBLEM SET 21.3

1–6 EULER FOR SYSTEMS AND SECOND-ORDER ODEs

Solve by the Euler's method. Graph the solution in the y_1y_2-plane. Calculate the errors.

1. $y_1' = 2y_1 - 4y_2, \quad y_2' = y_1 - 3y_2, \quad y_1(0) = 3,$ $y_2(0) = 0, \quad h = 0.1, \quad 10$ steps
2. **Spiral.** $y_1' = -y_1 + y_2, \quad y_2' = -y_1 - y_2, \quad y_1(0) = 0,$ $y_2(0) = 4, \quad h = 0.2, \quad 5$ steps
3. $y'' + \frac{1}{4}y = 0, \quad y(0) = 1, \quad y'(0) = 0, \quad h = 0.2,$ 5 steps
4. $y_1' = -3y_1 + y_2, \quad y_2' = y_1 - 3y_2, \quad y_1(0) = 2,$ $y_2(0) = 0, \quad h = 0.1, \quad 5$ steps
5. $y'' - y = x, \quad y(0) = 1, \quad y'(0) = -2, \quad h = 0.1,$ 5 steps
6. $y_1' = y_1, \quad y_2' = -y_2, \quad y_1(0) = 2, \quad y_2(0) = 2,$ $h = 0.1, \quad 10$ steps

7–10 RK FOR SYSTEMS

Solve by the classical RK.

7. The ODE in Prob. 5. By what factor did the error decrease?
8. The system in Prob. 2
9. The system in Prob. 1
10. The system in Prob. 4
11. **Pendulum equation** $y'' + \sin y = 0, \quad y(\pi) = 0,$ $y'(\pi) = 1,$ as a system, $h = 0.2$, 20 steps. How does your result fit into Fig. 93 in Sec. 4.5?
12. **Bessel Function J_0.** $xy'' + y' + xy = 0, \quad y(1) = 0.765198, \quad y'(1) = -0.440051, \quad h = 0.5, \quad 5$ steps. (This gives the standard solution $J_0(x)$ in Fig. 110 in Sec. 5.4.)
13. Verify the formulas and calculations for the Airy equation in Example 2 of the text.
14. **RKN.** The classical RK for a first-order ODE extends to second-order ODEs (E. J. Nyström, *Acta fenn.* No 13, 1925). If the ODE is $y'' = f(x, y)$, not containing y', then

$$k_1 = \tfrac{1}{2}hf(x_n, y_n)$$
$$k_2 = \tfrac{1}{2}hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}h(y_n' + \tfrac{1}{2}k_1)) = k_3$$
$$k_4 = \tfrac{1}{2}hf(x_n + h, y_n + h(y_n' + k_2))$$
$$y_{n+1} = y_n + h(y_n' + \tfrac{1}{3}(k_1 + 2k_2))$$
$$y_{n+1}' = y_n' + \tfrac{1}{8}(k_1 + 4k_2 + k_4).$$

Apply this RKN (Runge–Kutta–Nyström) method to the Airy ODE in Example 2 with $h = 0.2$ as before, to obtain approximate values of $\mathrm{Ai}(x)$.

15. **CAS EXPERIMENT. Backward Euler and Stiffness.** Extend Example 3 as follows.

(a) Verify the values in Table 21.13 and show them graphically as in Fig. 452.

(b) Compute and graph Euler values for h near the "critical" $h = 0.18$ to determine more exactly when instability starts.

(c) Compute and graph RK values for values of h between 0.2 and 0.3 to find h for which the RK approximation begins to increase away from the exact solution.

(d) Compute and graph backward Euler values for large h; confirm stability and investigate the error increase for growing h.

21.4 Methods for Elliptic PDEs

We have arrived at the second half of this chapter, which is devoted to numerics for partial differential equations (PDEs). As we have seen in Chap.12, there are many applications to PDEs, such as in dynamics, elasticity, heat transfer, electromagnetic theory, quantum mechanics, and others. Selected because of their importance in applications, the PDEs covered here include the Laplace equation, the Poisson equation, the heat equation, and the wave equation. By covering these equations based on their importance in applications we also selected equations that are important for theoretical considerations. Indeed, these equations serve as models for elliptic, parabolic, and hyperbolic PDEs. For example, the Laplace equation is a representative example of an elliptic type of PDE, and so forth.

Recall, from Sec. 12.4, that a PDE is called **quasilinear** if it is linear in the highest derivatives. Hence a second-order quasilinear PDE in two independent variables x, y is of the form

$$au_{xx} + 2bu_{xy} + cu_{yy} = F(x, y, u, u_x, u_y). \tag{1}$$

u is an unknown function of x and y (a solution sought). F is a given function of the indicated variables.

Depending on the discriminant $ac - b^2$, the PDE (1) is said to be of

elliptic type	if	$ac - b^2 > 0$	(example: *Laplace equation*)
parabolic type	if	$ac - b^2 = 0$	(example: *heat equation*)
hyperbolic type	if	$ac - b^2 < 0$	(example: *wave equation*).

Here, in the heat and wave equations, y is time t. The *coefficients* a, b, c may be functions of x, y, so that the type of (1) may be different in different regions of the xy-plane. This classification is not merely a formal matter but is of great practical importance because the general behavior of solutions differs from type to type and so do the additional conditions (boundary and initial conditions) that must be taken into account.

Applications involving ***elliptic equations*** usually lead to boundary value problems in a region R, called a *first boundary value problem* or **Dirichlet problem** if u is prescribed on the boundary curve C of R, a *second boundary value problem* or **Neumann problem** if $u_n = \partial u/\partial n$ (normal derivative of u) is prescribed on C, and a *third* or **mixed problem** if u is prescribed on a part of C and u_n on the remaining part. C usually is a closed curve (or sometimes consists of two or more such curves).

Difference Equations for the Laplace and Poisson Equations

In this section we develop numeric methods for the two most important elliptic PDEs that appear in applications. The two PDEs are the **Laplace equation**

$$\nabla^2 u = u_{xx} + u_{yy} = 0 \tag{2}$$

and the **Poisson equation**

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y). \tag{3}$$

The starting point for developing our numeric methods is the idea that we can replace the partial derivatives of these PDEs by corresponding **difference quotients**. Details are as follows:

To develop this idea, we start with the Taylor formula and obtain

$$\text{(4)} \quad \begin{aligned} &\text{(a)} \quad u(x + h, y) = u(x, y) + hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) + \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots \\ &\text{(b)} \quad u(x - h, y) = u(x, y) - hu_x(x, y) + \tfrac{1}{2}h^2 u_{xx}(x, y) - \tfrac{1}{6}h^3 u_{xxx}(x, y) + \cdots . \end{aligned}$$

We subtract (4b) from (4a), neglect terms in $h^3, h^4, \cdots$, and solve for u_x. Then

$$u_x(x, y) \approx \frac{1}{2h}[u(x+h, y) - u(x-h, y)]. \tag{5a}$$

Similarly,

$$u(x, y+k) = u(x, y) + ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots$$

and

$$u(x, y-k) = u(x, y) - ku_y(x, y) + \tfrac{1}{2}k^2 u_{yy}(x, y) + \cdots.$$

By subtracting, neglecting terms in $k^3, k^4, \cdots$, and solving for u_y we obtain

$$u_y(x, y) \approx \frac{1}{2k}[u(x, y+k) - u(x, y-k)]. \tag{5b}$$

We now turn to second derivatives. Adding (4a) and (4b) and neglecting terms in $h^4, h^5, \cdots$, we obtain $u(x+h, y) + u(x-h, y) \approx 2u(x, y) + h^2 u_{xx}(x, y)$. Solving for u_{xx} we have

$$u_{xx}(x, y) \approx \frac{1}{h^2}[u(x+h, y) - 2u(x, y) + u(x-h, y)]. \tag{6a}$$

Similarly,

$$u_{yy}(x, y) \approx \frac{1}{k^2}[u(x, y+k) - 2u(x, y) + u(x, y-k)]. \tag{6b}$$

We shall not need (see Prob. 1)

$$\begin{aligned} u_{xy}(x, y) \approx \frac{1}{4hk}[&u(x+h, y+k) - u(x-h, y+k) \\ &- u(x+h, y-k) + u(x-h, y-k)]. \end{aligned} \tag{6c}$$

Figure 453a shows the points $(x+h, y), (x-h, y), \cdots$ in (5) and (6).

We now substitute (6a) and (6b) into the ***Poisson equation*** (3), choosing $k = h$ to obtain a simple formula:

$$u(x+h, y) + u(x, y+h) + u(x-h, y) + u(x, y-h) - 4u(x, y) = h^2 f(x, y). \tag{7}$$

This is a **difference equation** corresponding to (3). Hence for the ***Laplace equation*** (2) the corresponding difference equation is

$$u(x+h, y) + u(x, y+h) + u(x-h, y) + u(x, y-h) - 4u(x, y) = 0. \tag{8}$$

h is called the **mesh size**. Equation (8) relates u at (x, y) to u at the four neighboring points shown in Fig. 453b. It has a remarkable interpretation: u at (x, y) equals the mean of the

values of u at the four neighboring points. This is an analog of the mean value property of harmonic functions (Sec. 18.6).

Those neighbors are often called E (East), N (North), W (West), S (South). Then Fig. 453b becomes Fig. 453c and (7) is

$$u(E) + u(N) + u(W) + u(S) - 4u(x, y) = h^2 f(x, y). \tag{7*}$$

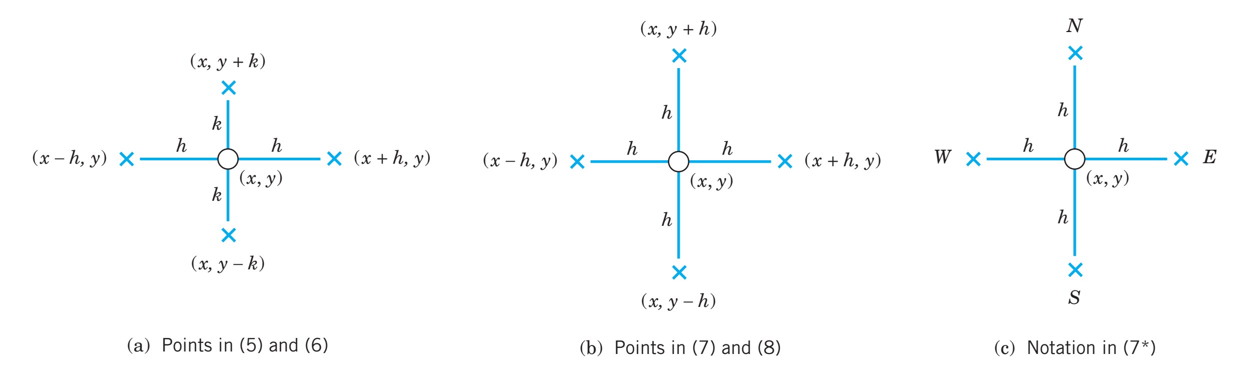


Fig. 453. Points and notation in (5)–(8) and (7*)

Our approximation of $h^2\nabla^2 u$ in (7) and (8) is a 5-point approximation with the coefficient scheme or **stencil** (also called *pattern, molecule,* or *star*)

$$\left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\}. \quad \text{We may now write (7) as} \quad \left\{\begin{matrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{matrix}\right\} u = h^2 f(x, y). \tag{9}$$

Dirichlet Problem

In numerics for the Dirichlet problem in a region R we choose an h and introduce a square grid of horizontal and vertical straight lines of distance h. Their intersections are called **mesh points** (or *lattice points* or *nodes*). See Fig. 454.

Then we approximate the given PDE by a difference equation [(8) for the Laplace equation], which relates the unknown values of u at the mesh points in R to each other and to the given boundary values (details in Example 1). This gives a linear system of *algebraic* equations. By solving it we get approximations of the unknown values of u at the mesh points in R.

We shall see that the number of equations equals the number of unknowns. Now comes an important point. If the number of internal mesh points, call it p, is small, say, $p < 100$, then a direct solution method may be applied to that linear system of $p < 100$ equations in p unknowns. However, if p is large, a storage problem will arise. Now since each unknown u is related to only 4 of its neighbors, the coefficient matrix of the system is a **sparse matrix**, that is, a matrix with relatively few nonzero entries (for instance, 500 of 10,000 when $p = 100$). Hence for large p we may avoid storage difficulties by using an iteration method, notably the Gauss–Seidel method (Sec. 20.3), which in PDEs is also

called **Liebmann's method** (note the strict diagonal dominance). Remember that in this method we have the storage convenience that we can overwrite any solution component (value of u) as soon as a "new" value is available.

Both cases, large p and small p, are of interest to the engineer, large p if a fine grid is used to achieve high accuracy, and small p if the boundary values are known only rather inaccurately, so that a coarse grid will do it because in this case it would be meaningless to try for great accuracy in the interior of the region R.

We illustrate this approach with an example, keeping the number of equations small, for simplicity. As convenient ***notations for mesh points and corresponding values of the solution*** (and of approximate solutions) we use (see also Fig. 454)

$$P_{ij} = (ih, jh), \qquad u_{ij} = u(ih, jh). \tag{10}$$

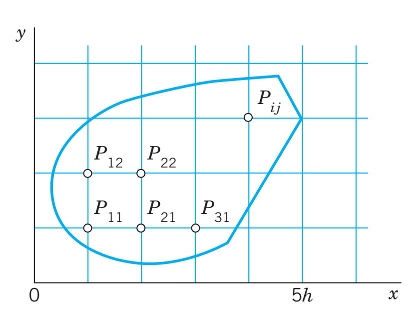


Fig. 454. Region in the xy-plane covered by a grid of mesh h, also showing mesh points $P_{11} = (h, h), \cdots, P_{ij} = (ih, jh), \cdots$

With this notation we can write (8) for any mesh point P_{ij} in the form

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0. \tag{11}$$

Remark. Our current discussion and the example that follows illustrate what we may call the ***reuseability of mathematical ideas and methods***. Recall that we applied the Gauss–Seidel method to a system of ODEs in Sec. 20.3 and that we can now apply it again to elliptic PDEs. This shows that engineering mathematics has a structure and important mathematical ideas and methods will appear again and again in different situations. The student should find this attractive in that previous knowledge can be reapplied.

EXAMPLE 1 Laplace Equation. Liebmann's Method

The four sides of a square plate of side 12 cm, made of homogeneous material, are kept at constant temperature 0°C and 100°C as shown in Fig. 455a. Using a (very wide) grid of mesh 4 cm and applying Liebmann's method (that is, Gauss–Seidel iteration), find the (steady-state) temperature at the mesh points.

Solution. In the case of independence of time, the heat equation (see Sec. 10.8)

$$u_t = c^2(u_{xx} + u_{yy})$$

reduces to the Laplace equation. Hence our problem is a Dirichlet problem for the latter. We choose the grid shown in Fig. 455b and consider the mesh points in the order $P_{11}, P_{21}, P_{12}, P_{22}$. We use (11) and, in each equation, take to the right all the terms resulting from the given boundary values. Then we obtain the system

$$
\begin{aligned}
-4u_{11} + u_{21} + u_{12} \qquad\qquad &= -200 \\
-u_{11} - 4u_{21} \qquad + u_{22} &= -200 \\
u_{11} \qquad - 4u_{12} + u_{22} &= -100 \\
u_{21} + u_{12} - 4u_{22} &= -100.
\end{aligned} \tag{12}
$$

In practice, one would solve such a small system by the Gauss elimination, finding $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$.

More exact values (exact to 3S) of the solution of the actual problem [as opposed to its model (12)] are 88.1 and 61.9, respectively. (These were obtained by using Fourier series.) Hence the error is about 1%, which is surprisingly accurate for a grid of such a large mesh size h. If the system of equations were large, one would solve it by an indirect method, such as Liebmann's method. For (12) this is as follows. We write (12) in the form (divide by -4 and take terms to the right)

$$
\begin{aligned}
u_{11} &= \qquad\quad 0.25u_{21} + 0.25u_{12} \qquad\qquad + 50 \\
u_{21} &= 0.25u_{11} \qquad\qquad\qquad + 0.25u_{22} + 50 \\
u_{12} &= 0.25u_{11} \qquad\qquad\qquad + 0.25u_{22} + 25 \\
u_{22} &= \qquad\quad 0.25u_{21} + 0.25u_{12} \qquad\qquad + 25.
\end{aligned}
$$

These equations are now used for the Gauss–Seidel iteration. They are identical with (2) in Sec. 20.3, where $u_{11} = x_1$, $u_{21} = x_2$, $u_{12} = x_3$, $u_{22} = x_4$, and the iteration is explained there, with 100, 100, 100, 100 chosen as starting values. Some work can be saved by better starting values, usually by taking the average of the boundary values that enter into the linear system. The exact solution of the system is $u_{11} = u_{21} = 87.5$, $u_{12} = u_{22} = 62.5$, as you may verify.

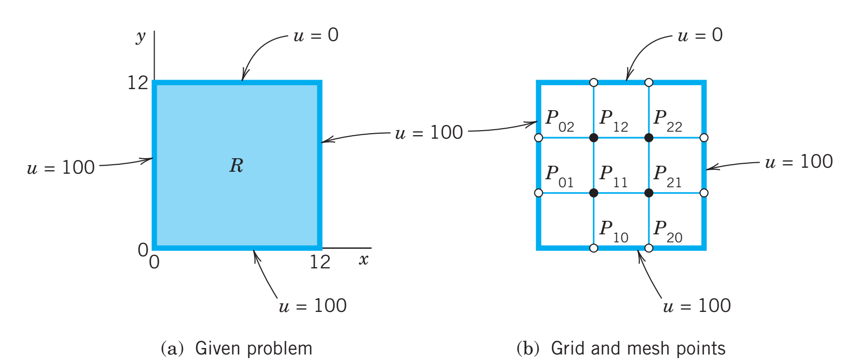


Fig. 455. Example 1

Remark. It is interesting to note that, if we choose mesh $h = L/n$ (L = side of R) and consider the $(n-1)^2$ internal mesh points (i.e., mesh points not on the boundary) row by row in the order

$$P_{11}, P_{21}, \cdots, P_{n-1,1}, P_{12}, P_{22}, \cdots, P_{n-2,2}, \cdots,$$

then the system of equations has the $(n-1)^2 \times (n-1)^2$ coefficient matrix

$$
\mathbf{A} = \begin{bmatrix} \mathbf{B} & \mathbf{I} & & & & \\ \mathbf{I} & \mathbf{B} & \mathbf{I} & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & \mathbf{I} & \mathbf{B} & \mathbf{I} \\ & & & & \mathbf{I} & \mathbf{B} \end{bmatrix}. \qquad \text{Here} \qquad \mathbf{B} = \begin{bmatrix} -4 & 1 & & & & \\ 1 & -4 & 1 & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & 1 & -4 & 1 \\ & & & & 1 & -4 \end{bmatrix} \tag{13}
$$

is an $(n-1) \times (n-1)$ matrix. (In (12) we have $n = 3$, $(n-1)^2 = 4$ internal mesh points, two submatrices **B**, and two submatrices **I**.) The matrix **A** is nonsingular. This follows by noting that the off-diagonal entries in each row of **A** have the sum 3 (or 2), whereas each diagonal entry of **A** equals -4, so that nonsingularity is implied by Gerschgorin's theorem in Sec. 20.7 because no Gerschgorin disk can include 0. ■

A matrix is called a **band matrix** if it has all its nonzero entries on the main diagonal and on sloping lines parallel to it (separated by sloping lines of zeros or not). For example, **A** in (13) is a band matrix. Although the Gauss elimination does not preserve zeros between bands, it does not introduce nonzero entries outside the limits defined by the original bands. Hence a band structure is advantageous. In (13) it has been achieved by carefully ordering the mesh points.

ADI Method

A matrix is called a **tridiagonal matrix** if it has all its nonzero entries on the main diagonal and on the two sloping parallels immediately above or below the diagonal. (See also Sec. 20.9.) In this case the Gauss elimination is particularly simple.

This raises the question of whether, in the solution of the Dirichlet problem for the Laplace or Poisson equations, one could obtain a system of equations whose coefficient matrix is tridiagonal. The answer is yes, and a popular method of that kind, called the **ADI method** (*alternating direction implicit method*) was developed by Peaceman and Rachford. The idea is as follows. The stencil in (9) shows that we could obtain a tridiagonal matrix if there were only the three points in a row (or only the three points in a column). This suggests that we write (11) in the form

$$u_{i-1,j} - 4u_{ij} + u_{i+1,j} = -u_{i,j-1} - u_{i,j+1} \tag{14a}$$

so that the left side belongs to y-Row j only and the right side to x-Column i. Of course, we can also write (11) in the form

$$u_{i,j-1} - 4u_{ij} + u_{i,j+1} = -u_{i-1,j} - u_{i+1,j} \tag{14b}$$

so that the left side belongs to Column i and the right side to Row j. In the ADI method we proceed by iteration. At every mesh point we choose an arbitrary starting value $u_{ij}^{(0)}$. In each step we compute new values at all mesh points. In one step we use an iteration formula resulting from (14a) and in the next step an iteration formula resulting from (14b), and so on in alternating order.

In detail: suppose approximations $u_{ij}^{(m)}$ have been computed. Then, to obtain the next approximations $u_{ij}^{(m+1)}$, we substitute the $u_{ij}^{(m)}$ ***on the right*** side of (14a) and solve for the $u_{ij}^{(m+1)}$ on the left side; that is, we use

$$u_{i-1,j}^{(m+1)} - 4u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} - u_{i,j+1}^{(m)}. \tag{15a}$$

We use (15a) for a fixed j, that is, ***for a fixed row*** j, and for all internal mesh points in this row. This gives a linear system of N algebraic equations (N = number of internal mesh points per row) in N unknowns, the new approximations of u at these mesh points. Note that (15a) involves not only approximations computed in the previous step but also given boundary values. We solve the system (15a) (j fixed!) by Gauss elimination. Then we go to the next row, obtain another system of N equations and solve it by Gauss, and so on, until all rows are done. In the next step we ***alternate direction***, that is, we compute

the next approximations $u_{ij}^{(m+2)}$ column by column from the $u_{ij}^{(m+1)}$ and the given boundary values, using a formula obtained from (14b) by substituting the $u_{ij}^{(m+1)}$ ***on the right***:

(15b) $$u_{i,j-1}^{(m+2)} - 4u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} - u_{i+1,j}^{(m+1)}.$$

For each fixed i, that is, ***for each column,*** this is a system of M equations (M = number of internal mesh points per column) in M unknowns, which we solve by Gauss elimination. Then we go to the next column, and so on, until all columns are done.

Let us consider an example that merely serves to explain the entire method.

EXAMPLE 2 Dirichlet Problem. ADI Method

Explain the procedure and formulas of the ADI method in terms of the problem in Example 1, using the same grid and starting values 100, 100, 100, 100.

Solution. While working, we keep an eye on Fig. 455b and the given boundary values. We obtain first approximations $u_{11}^{(1)}, u_{21}^{(1)}, u_{12}^{(1)}, u_{22}^{(1)}$ from (15a) with $m = 0$. We write boundary values contained in (15a) without an upper index, for better identification and to indicate that these given values remain the same during the iteration. From (15a) with $m = 0$ we have for $j = 1$ (first row) the system

$$\begin{aligned} &(i=1) \quad u_{01} - 4u_{11}^{(1)} + u_{21}^{(1)} &&= -u_{10} - u_{12}^{(0)} \\ &(i=2) \quad u_{11}^{(1)} - 4u_{21}^{(1)} + u_{31} &&= -u_{20} - u_{22}^{(0)}. \end{aligned}$$

The solution is $u_{11}^{(1)} = u_{21}^{(1)} = 100$. For $j = 2$ (second row) we obtain from (15a) the system

$$\begin{aligned} &(i=1) \quad u_{02} - 4u_{12}^{(1)} + u_{22}^{(1)} &&= -u_{11}^{(0)} - u_{13} \\ &(i=2) \quad u_{12}^{(1)} - 4u_{22}^{(1)} + u_{32} &&= -u_{21}^{(0)} - u_{23}. \end{aligned}$$

The solution is $u_{12}^{(1)} = u_{22}^{(1)} = 66.667$.

Second approximations $u_{11}^{(2)}, u_{21}^{(2)}, u_{12}^{(2)}, u_{22}^{(2)}$ are now obtained from (15b) with $m = 1$ by using the first approximations just computed and the boundary values. For $i = 1$ (first column) we obtain from (15b) the system

$$\begin{aligned} &(j=1) \quad u_{10} - 4u_{11}^{(2)} + u_{12}^{(2)} &&= -u_{01} - u_{21}^{(1)} \\ &(j=2) \quad u_{11}^{(2)} - 4u_{12}^{(2)} + u_{13} &&= -u_{02} - u_{22}^{(1)}. \end{aligned}$$

The solution is $u_{11}^{(2)} = 91.11$, $u_{12}^{(2)} = 64.44$, For $i = 2$ (second column) we obtain from (15b) the system

$$\begin{aligned} &(j=1) \quad u_{20} - 4u_{21}^{(2)} + u_{22}^{(2)} &&= -u_{11}^{(1)} - u_{31} \\ &(j=2) \quad u_{21}^{(2)} - 4u_{22}^{(2)} + u_{23} &&= -u_{12}^{(1)} - u_{32}. \end{aligned}$$

The solution is $u_{21}^{(2)} = 91.11$, $u_{22}^{(2)} = 64.44$.

In this example, which merely serves to explain the practical procedure in the ADI method, the accuracy of the second approximations is about the same as that of two Gauss–Seidel steps in Sec. 20.3 (where $u_{11} = x_1, u_{21} = x_2, u_{12} = x_3, u_{22} = x_4$), as the following table shows.

Method	u_{11}	u_{21}	u_{12}	u_{22}
ADI, 2nd approximations	91.11	91.11	64.44	64.44
Gauss–Seidel, 2nd approximations	93.75	90.62	65.62	64.06
Exact solution of (12)	87.50	87.50	62.50	62.50

■

Improving Convergence. Additional improvement of the convergence of the ADI method results from the following interesting idea. Introducing a parameter p, we can also write (11) in the form

$$\text{(16)} \quad \begin{aligned} &\text{(a)} \quad u_{i-1,j} - (2+p)u_{ij} + u_{i+1,j} = -u_{i,j-1} + (2-p)u_{ij} - u_{i,j+1} \\ &\text{(b)} \quad u_{i,j-1} - (2+p)u_{ij} + u_{i,j+1} = -u_{i-1,j} + (2-p)u_{ij} - u_{i+1,j}. \end{aligned}$$

This gives the more general ADI iteration formulas

$$\text{(17)} \quad \begin{aligned} &\text{(a)} \quad u_{i-1,j}^{(m+1)} - (2+p)u_{ij}^{(m+1)} + u_{i+1,j}^{(m+1)} = -u_{i,j-1}^{(m)} + (2-p)u_{ij}^{(m)} - u_{i,j+1}^{(m)} \\ &\text{(b)} \quad u_{i,j-1}^{(m+2)} - (2+p)u_{ij}^{(m+2)} + u_{i,j+1}^{(m+2)} = -u_{i-1,j}^{(m+1)} + (2-p)u_{ij}^{(m+1)} - u_{i+1,j}^{(m+1)}. \end{aligned}$$

For $p = 2$, this is (15). The parameter p may be used for improving convergence. Indeed, one can show that the ADI method converges for positive p, and that the optimum value for maximum rate of convergence is

$$\text{(18)} \qquad p_0 = 2 \sin \frac{\pi}{K}$$

where K is the larger of $M + 1$ and $N + 1$ (see above). Even better results can be achieved by letting p vary from step to step. More details of the ADI method and variants are discussed in Ref. [E25] listed in App. 1.

PROBLEM SET 21.4

1. Derive (5b), (6b), and (6c).
2. Verify the calculations in Example 1 of the text. Find out experimentally how many steps you need to obtain the solution of the linear system with an accuracy of 3S.
3. **Use of symmetry.** Conclude from the boundary values in Example 1 that $u_{21} = u_{11}$ and $u_{22} = u_{12}$. Show that this leads to a system of two equations and solve it.
4. **Finer grid** of 3×3 inner points. Solve Example 1, choosing $h = \frac{12}{4} = 3$ (instead of $h = \frac{12}{3} = 4$) and the same starting values.

5–10 GAUSS ELIMINATION, GAUSS–SEIDEL ITERATION

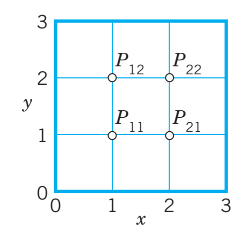


Fig. 456. Problems 5–10

For the grid in Fig. 456 compute the potential at the four internal points by Gauss and by 5 Gauss–Seidel steps with starting values 100, 100, 100, 100 (showing the details of your work) if the boundary values on the edges are:

5. $u(1, 0) = 60$, $u(2, 0) = 300$, $u = 100$ on the other three edges.
6. $u = 0$ on the left, x^3 on the lower edge, $27 - 9y^2$ on the right, $x^3 - 27x$ on the upper edge.
7. U_0 on the upper and lower edges, $-U_0$ on the left and right. Sketch the equipotential lines.
8. $u = 220$ on the upper and lower edges, 110 on the left and right.
9. $u = \sin \frac{1}{3}\pi x$ on the upper edge, 0 on the other edges, 10 steps.
10. $u = x^4$ on the lower edge, $81 - 54y^2 + y^4$ on the right, $x^4 - 54x^2 + 81$ on the upper edge, y^4 on the left. Verify the exact solution $x^4 - 6x^2y^2 + y^4$ and determine the error.

11. Find the potential in Fig. 457 using **(a)** the coarse grid, **(b)** the fine grid 5×3, and Gauss elimination. *Hint.* In (b), use symmetry; take $u = 0$ as boundary value at the two points at which the potential has a jump.

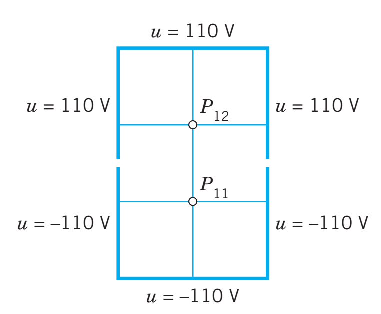


Fig. 457. Region and grids in Problem 11

12. Influence of starting values. Do Prob. 9 by Gauss–Seidel, starting from **0**. Compare and comment.

13. For the square $0 \leqq x \leqq 4$, $0 \leqq y \leqq 4$ let the boundary temperatures be 0°C on the horizontal and 50°C on the vertical edges. Find the temperatures at the interior points of a square grid with $h = 1$.

14. Using the answer to Prob. 13, try to sketch some isotherms.

15. Find the isotherms for the square and grid in Prob. 13 if $u = \sin \frac{1}{4}\pi x$ on the horizontal and $-\sin \frac{1}{4}\pi y$ on the vertical edges. Try to sketch some isotherms.

16. ADI. Apply the ADI method to the Dirichlet problem in Prob. 9, using the grid in Fig. 456, as before and starting values zero.

17. What p_0 in (18) should we choose for Prob. 16? Apply the ADI formulas (17) with that value of p_0 to Prob. 16, performing 1 step. Illustrate the improved convergence by comparing with the corresponding values 0.077, 0.308 after the first step in Prob. 16. (Use the starting values zero.)

18. CAS PROJECT. Laplace Equation. (a) Write a program for Gauss–Seidel with 16 equations in 16 unknowns, composing the matrix (13) from the indicated 4×4 submatrices and including a transformation of the vector of the boundary values into the vector $\mathbf{b}$ of $\mathbf{Ax} = \mathbf{b}$.

(b) Apply the program to the square grid in $0 \leqq x \leqq 5$, $0 \leqq y \leqq 5$ with $h = 1$ and $u = 220$ on the upper and lower edges, $u = 110$ on the left edge and $u = -10$ on the right edge. Solve the linear system also by Gauss elimination. What accuracy is reached in the 20th Gauss–Seidel step?

21.5 Neumann and Mixed Problems. Irregular Boundary

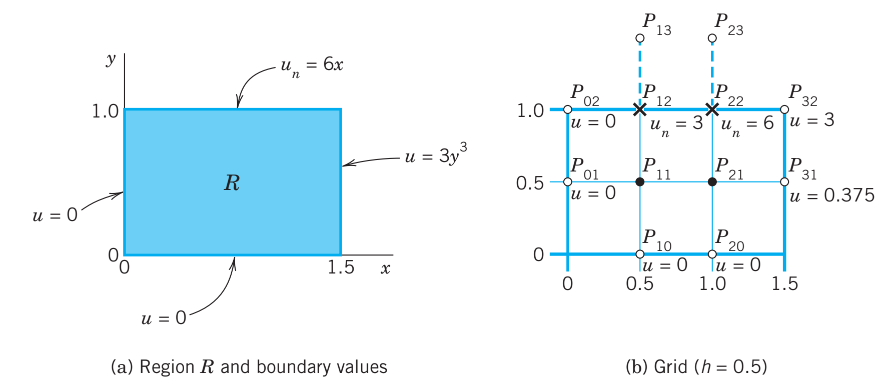
We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. The Dirichlet problem was studied in the last section. In solving **Neumann** and **mixed problems** (defined in the last section) we are confronted with a new situation, because there are boundary points at which the (outer) **normal derivative** $u_n = \partial u/\partial n$ of the solution is given, but u itself is unknown since it is not given. To handle such points we need a new idea. This idea is the same for Neumann and mixed problems. Hence we may explain it in connection with one of these two types of problems. We shall do so and consider a typical example as follows.

EXAMPLE 1 Mixed Boundary Value Problem for a Poisson Equation

Solve the mixed boundary value problem for the Poisson equation

$$\nabla^2 u = u_{xx} + u_{yy} = f(x, y) = 12xy$$

shown in Fig. 458a.



(a) Region R and boundary values (b) Grid ($h = 0.5$)

Fig. 458. Mixed boundary value problem in Example 1

Solution. We use the grid shown in Fig. 458b, where $h = 0.5$. We recall that (7) in Sec. 21.4 has the right side $h^2 f(x, y) = 0.5^2 \cdot 12xy = 3xy$. From the formulas $u = 3y^3$ and $u_n = 6x$ given on the boundary we compute the boundary data

$$u_{31} = 0.375, \qquad u_{32} = 3, \qquad \frac{\partial u_{12}}{\partial n} = \frac{\partial u_{12}}{\partial y} = 6 \cdot 0.5 = 3, \qquad \frac{\partial u_{22}}{\partial n} = \frac{\partial u_{22}}{\partial y} = 6 \cdot 1 = 6. \tag{1}$$

P_{11} and P_{21} are internal mesh points and can be handled as in the last section. Indeed, from (7), Sec. 21.4, with $h^2 = 0.25$ and $h^2 f(x, y) = 3xy$ and from the given boundary values we obtain two equations corresponding to P_{11} and P_{21}, as follows (with -0 resulting from the left boundary).

$$\begin{aligned} -4u_{11} + u_{21} + u_{12} \qquad\qquad &= 12(0.5 \cdot 0.5) \cdot \tfrac{1}{4} - 0 = 0.75 \\ u_{11} - 4u_{21} \qquad + u_{22} &= 12(1 \cdot 0.5) \cdot \tfrac{1}{4} - 0.375 = 1.125. \end{aligned} \tag{2a}$$

The only difficulty with these equations seems to be that they involve the unknown values u_{12} and u_{22} of u at P_{12} and P_{22} on the boundary, where the normal derivative $u_n = \partial u/\partial n = \partial u/\partial y$ is given, instead of u; but we shall overcome this difficulty as follows.

We consider P_{12} and P_{22}. The idea that will help us here is this. We imagine the region R to be extended above to the first row of external mesh points (corresponding to $y = 1.5$), and we assume that the Poisson equation also holds in the extended region. Then we can write down two more equations as before (Fig. 458b)

$$\begin{aligned} u_{11} \qquad - 4u_{12} + u_{22} + u_{13} \qquad &= 1.5 - 0 = 1.5 \\ u_{21} + u_{12} - 4u_{22} \qquad + u_{23} &= 3 - 3 = 0. \end{aligned} \tag{2b}$$

On the right, 1.5 is $12xyh^2$ at (0.5, 1) and 3 is $12xyh^2$ at (1, 1) and 0 (at P_{02}) and 3 (at P_{32}) are given boundary values. We remember that we have not yet used the boundary condition on the upper part of the boundary of R, and we also notice that in (2b) we have introduced two more unknowns u_{13}, u_{23}. But we can now use that condition and get rid of u_{13}, u_{23} by applying the central difference formula for du/dy. From (1) we then obtain (see Fig. 458b)

$$\begin{aligned} 3 &= \frac{\partial u_{12}}{\partial y} \approx \frac{u_{13} - u_{11}}{2h} = u_{13} - u_{11}, \qquad \text{hence} \qquad u_{13} = u_{11} + 3 \\ 6 &= \frac{\partial u_{22}}{\partial y} \approx \frac{u_{23} - u_{21}}{2h} = u_{23} - u_{21}, \qquad \text{hence} \qquad u_{23} = u_{21} + 6. \end{aligned}$$

Substituting these results into (2b) and simplifying, we have

$$\begin{aligned} 2u_{11} \qquad - 4u_{12} + u_{22} &= 1.5 - 3 = -1.5 \\ 2u_{21} + u_{12} - 4u_{22} &= 3 - 3 - 6 = -6. \end{aligned}$$

Together with (2a) this yields, written in matrix form,

$$(3) \qquad \begin{bmatrix} -4 & 1 & 1 & 0 \\ 1 & -4 & 0 & 1 \\ 2 & 0 & -4 & 1 \\ 0 & 2 & 1 & -4 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ 1.5 - 3 \\ 0 - 6 \end{bmatrix} = \begin{bmatrix} 0.75 \\ 1.125 \\ -1.5 \\ -6 \end{bmatrix}.$$

(The entries 2 come from u_{13} and u_{23}, and so do -3 and -6 on the right). The solution of (3) (obtained by Gauss elimination) is as follows; the exact values of the problem are given in parentheses.

$$u_{12} = 0.866 \quad \text{(exact 1)} \qquad u_{22} = 1.812 \quad \text{(exact 2)}$$

$$u_{11} = 0.077 \quad \text{(exact 0.125)} \qquad u_{21} = 0.191 \quad \text{(exact 0.25)}.$$

■

Irregular Boundary

We continue our discussion of boundary value problems for elliptic PDEs in a region R in the xy-plane. If R has a simple geometric shape, we can usually arrange for certain mesh points to lie on the boundary C of R, and then we can approximate partial derivatives as explained in the last section. However, if C intersects the grid at points that are not mesh points, then at points close to the boundary we must proceed differently, as follows.

The mesh point O in Fig. 459 is of that kind. For O and its neighbors A and P we obtain from Taylor's theorem

$$(4) \qquad \begin{aligned} &\text{(a)} \quad u_A = u_O + ah\frac{\partial u_O}{\partial x} + \frac{1}{2}(ah)^2 \frac{\partial^2 u_O}{\partial x^2} + \cdots \\ &\text{(b)} \quad u_P = u_O - h\frac{\partial u_O}{\partial x} + \frac{1}{2}h^2 \frac{\partial^2 u_O}{\partial x^2} + \cdots . \end{aligned}$$

We disregard the terms marked by dots and eliminate $\partial u_O / \partial x$. Equation (4b) times a plus equation (4a) gives

$$u_A + au_P \approx (1 + a)\, u_O + \frac{1}{2}\, a\,(a + 1)\, h^2 \frac{\partial^2 u_O}{\partial x^2}.$$

Fig. 459. Curved boundary C of a region R, a mesh point O near C, and neighbors A, B, P, Q

We solve this last equation algebraically for the derivative, obtaining

$$\frac{\partial^2 u_O}{\partial x^2} \approx \frac{2}{h^2}\left[\frac{1}{a(1 + a)}\, u_A + \frac{1}{1 + a}\, u_P - \frac{1}{a}\, u_O\right].$$

Similarly, by considering the points O, B, and Q,

$$\frac{\partial^2 u_O}{\partial y^2} \approx \frac{2}{h^2}\left[\frac{1}{b(1+b)}u_B + \frac{1}{1+b}u_Q - \frac{1}{b}u_O\right].$$

By addition,

$$\text{(5)} \qquad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(1+a)} + \frac{u_B}{b(1+b)} + \frac{u_P}{1+a} + \frac{u_Q}{1+b} - \frac{(a+b)u_O}{ab}\right].$$

For example, if $a = \frac{1}{2}$, $b = \frac{1}{2}$, instead of the stencil (see Sec. 21.4)

$$\begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix} \qquad \text{we now have} \qquad \begin{Bmatrix} & \frac{4}{3} & \\ \frac{2}{3} & -4 & \frac{4}{3} \\ & \frac{2}{3} & \end{Bmatrix}.$$

because $1/[a(1+a)] = \frac{4}{3}$, etc. The sum of all five terms still being zero (which is useful for checking).

Using the same ideas, you may show that in the case of Fig. 460.

$$\textbf{(6)} \qquad \nabla^2 u_O \approx \frac{2}{h^2}\left[\frac{u_A}{a(a+p)} + \frac{u_B}{b(b+q)} + \frac{u_P}{p(p+a)} + \frac{u_Q}{q(q+b)} - \frac{ap+bq}{abpq}u_O\right],$$

a formula that takes care of all conceivable cases.

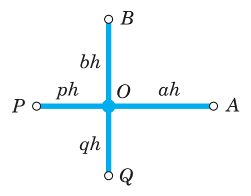


Fig. 460. Neighboring points A, B, P, Q of a mesh point O and notations in formula (6)

EXAMPLE 2 Dirichlet Problem for the Laplace Equation. Curved Boundary

Find the potential u in the region in Fig. 461 that has the boundary values given in that figure; here the curved portion of the boundary is an arc of the circle of radius 10 about (0,0). Use the grid in the figure.

Solution. u is a solution of the Laplace equation. From the given formulas for the boundary values $u = x^3$, $u = 512 - 24y^2, \cdots$ we compute the values at the points where we need them; the result is shown in the figure. For P_{11} and P_{12} we have the usual regular stencil, and for P_{21} and P_{22} we use (6), obtaining

$$\text{(7)} \qquad P_{11}, P_{12}: \begin{Bmatrix} & 1 & \\ 1 & -4 & 1 \\ & 1 & \end{Bmatrix}, \qquad P_{21}: \begin{Bmatrix} & 0.5 & \\ 0.6 & -2.5 & 0.9 \\ & 0.5 & \end{Bmatrix}, \qquad P_{22}: \begin{Bmatrix} & 0.9 & \\ 0.6 & -3 & 0.9 \\ & 0.6 & \end{Bmatrix}.$$

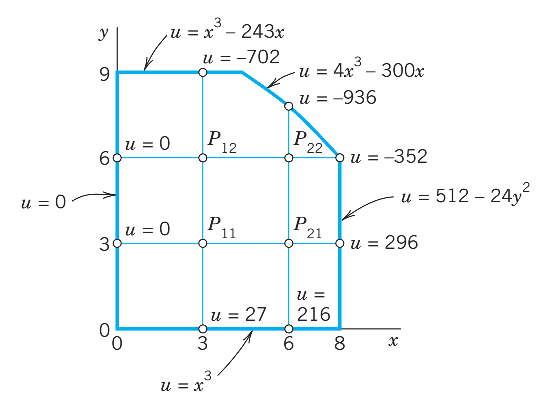


Fig. 461. Region, boundary values of the potential, and grid in Example 2

We use this and the boundary values and take the mesh points in the usual order $P_{11}, P_{21}, P_{12}, P_{22}$. Then we obtain the system

$$\begin{aligned}
-4u_{11} + u_{21} + u_{12} \qquad\qquad &= 0 - 27 &&= -27 \\
0.6u_{11} - 2.5u_{21} \qquad + 0.5u_{22} &= -0.9 \cdot 296 - 0.5 \cdot 216 &&= -374.4 \\
u_{11} \qquad - 4u_{12} + u_{22} &= 702 + 0 &&= 702 \\
0.6u_{21} + 0.6u_{12} - 3u_{22} &= 0.9 \cdot 352 + 0.9 \cdot 936 &&= 1159.2
\end{aligned}$$

In matrix form,

$$\begin{bmatrix} -4 & 1 & 1 & 0 \\ 0.6 & -2.5 & 0 & 0.5 \\ 1 & 0 & -4 & 1 \\ 0 & 0.6 & 0.6 & -3 \end{bmatrix} \begin{bmatrix} u_{11} \\ u_{21} \\ u_{12} \\ u_{22} \end{bmatrix} = \begin{bmatrix} -27 \\ -374.4 \\ 702 \\ 1159.2 \end{bmatrix}. \tag{8}$$

Gauss elimination yields the (rounded) values

$$u_{11} = -55.6, \qquad u_{21} = 49.2, \qquad u_{12} = -298.5, \qquad u_{22} = -436.3.$$

Clearly, from a grid with so few mesh points we cannot expect great accuracy. The exact solution of the PDE (not of the difference equation) having the given boundary values is $u = x^3 - 3xy^2$ and yields the values

$$u_{11} = -54, \qquad u_{21} = 54, \qquad u_{12} = -297, \qquad u_{22} = -432.$$

In practice one would use a much finer grid and solve the resulting large system by an indirect method. ■

PROBLEM SET 21.5

1–7 MIXED BOUNDARY VALUE PROBLEMS

1. Check the values for the Poisson equation at the end of Example 1 by solving (3) by Gauss elimination.

2. Solve the mixed boundary value problem for the Poisson equation $\nabla^2 u = 2(x^2 + y^2)$ in the region and for the boundary conditions shown in Fig. 462, using the indicated grid.

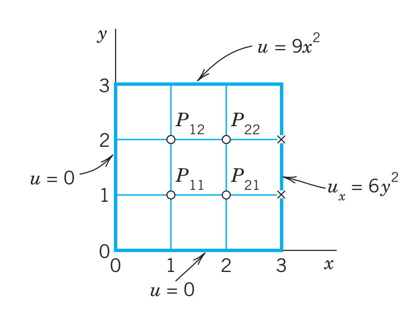


Fig. 462. Problems 2 and 6

3. CAS EXPERIMENT. Mixed Problem. Do Example 1 in the text with finer and finer grids of your choice and study the accuracy of the approximate values by comparing with the exact solution $u = 2xy^3$. Verify the latter.

4. Solve the mixed boundary value problem for the Laplace equation $\nabla^2 u = 0$ in the rectangle in Fig. 458a (using the grid in Fig. 458b) and the boundary conditions $u_x = 0$ on the left edge, $u_x = 3$ on the right edge, $u = x^2$ on the lower edge, and $u = x^2 - 1$ on the upper edge.

5. Do Example 1 in the text for the Laplace equation (instead of the Poisson equation) with grid and boundary data as before.

6. Solve $\nabla^2 u = -\pi^2 y \sin \frac{1}{3}\pi x$ for the grid in Fig. 462 and $u_y(1, 3) = u_y(2, 3) = \frac{1}{2}\sqrt{243}$, $u = 0$ on the other three sides of the square.

7. Solve Prob. 4 when $u_n = 110$ on the upper edge and $u = 110$ on the other edges.

8–16 IRREGULAR BOUNDARY

8. Verify the stencil shown after (5).

9. Derive (5) in the general case.

10. Derive the general formula (6) in detail.

11. Derive the linear system in Example 2 of the text.

12. Verify the solution in Example 2.

13. Solve the Laplace equation in the region and for the boundary values shown in Fig. 463, using the indicated grid. (The sloping portion of the boundary is $y = 4.5 - x$.)

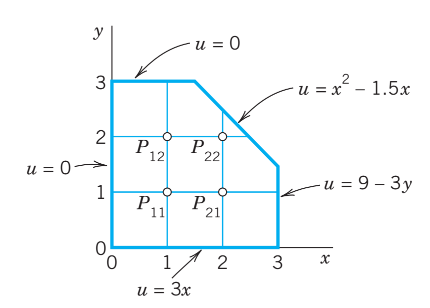


Fig. 463. Problem 13

14. If, in Prob. 13, the axes are grounded ($u = 0$), what constant potential must the other portion of the boundary have in order to produce 220 V at P_{11}?

15. What potential do we have in Prob. 13 if $u = 100$ V on the axes and $u = 0$ on the other portion of the boundary?

16. Solve the Poisson equation $\nabla^2 u = 2$ in the region and for the boundary values shown in Fig. 464, using the grid also shown in the figure.

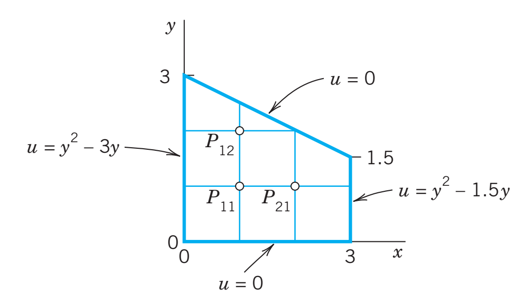


Fig. 464. Problem 16

21.6 Methods for Parabolic PDEs

The last two sections concerned elliptic PDEs, and we now turn to parabolic PDEs. Recall that the definitions of elliptic, parabolic, and hyperbolic PDEs were given in Sec. 21.4. There it was also mentioned that the general behavior of solutions differs from type to type, and so do the problems of practical interest. This reflects on numerics as follows.

For all three types, one replaces the PDE by a corresponding difference equation, but for *parabolic* and *hyperbolic* PDEs this does not automatically guarantee the **convergence** of the approximate solution to the exact solution as the mesh $h \to 0$; in fact, it does not even guarantee convergence at all. For these two types of PDEs one needs additional conditions (inequalities) to assure convergence and **stability**, the latter meaning that small perturbations in the initial data (or small errors at any time) cause only small changes at later times.

In this section we explain the numeric solution of the prototype of parabolic PDEs, the one-dimensional heat equation

$$u_t = c^2 u_{xx} \qquad (c \text{ constant}).$$

This PDE is usually considered for x in some fixed interval, say, $0 \leqq x \leqq L$, and time $t \geqq 0$, and one prescribes the initial temperature $u(x, 0) = f(x)$ (f given) and boundary conditions at $x = 0$ and $x = L$ for all $t \geqq 0$, for instance, $u(0, t) = 0$, $u(L, t) = 0$. We may assume $c = 1$ and $L = 1$; this can always be accomplished by a linear transformation of x and t (Prob. 1). Then the **heat equation** and those conditions are

(1) $$u_t = u_{xx} \qquad 0 \leqq x \leqq 1,\ t \geqq 0$$

(2) $$u(x, 0) = f(x) \qquad \text{(Initial condition)}$$

(3) $$u(0, t) = u(1, t) = 0 \qquad \text{(Boundary conditions).}$$

A simple finite difference approximation of (1) is [see (6a) in Sec. 21.4; j is the number of the ***time step***]

(4) $$\frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}).$$

Figure 465 shows a corresponding grid and mesh points. The mesh size is h in the x-direction and k in the t-direction. Formula (4) involves the four points shown in Fig. 466. On the left in (4) we have used a *forward* difference quotient since we have no information for negative t at the start. From (4) we calculate $u_{i,j+1}$, which corresponds to time row $j + 1$, in terms of the three other u that correspond to time row j. Solving (4) for $u_{i,j+1}$, we have

(5) $$u_{i,j+1} = (1 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}), \qquad r = \frac{k}{h^2}.$$

Computations by this **explicit method** based on (5) are simple. However, it can be shown that crucial to the convergence of this method is the condition

(6) $$r = \frac{k}{h^2} \leqq \frac{1}{2}.$$

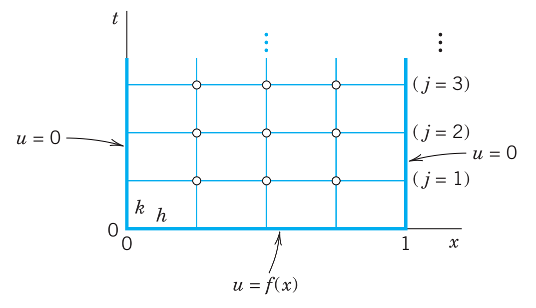


Fig. 465. Grid and mesh points corresponding to (4), (5)

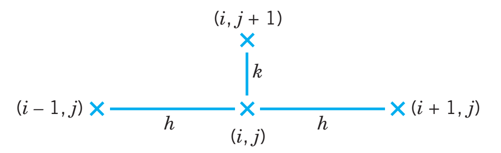


Fig. 466. The four points in (4) and (5)

That is, u_{ij} should have a positive coefficient in (5) or (for $r = \frac{1}{2}$) be absent from (5). Intuitively, (6) means that we should not move too fast in the t-direction. An example is given below.

Crank–Nicolson Method

Condition (6) is a handicap in practice. Indeed, to attain sufficient accuracy, we have to choose h small, which makes k very small by (6). For example, if $h = 0.1$, then $k \leqq 0.005$. Accordingly, we should look for a more satisfactory discretization of the heat equation.

A method that imposes no restriction on $r = k/h^2$ is the **Crank–Nicolson (CN) method**,[5] which uses values of u at the six points in Fig. 467. The idea of the method is the replacement of the difference quotient on the right side of (4) by $\frac{1}{2}$ times the sum of two such difference quotients at two time rows (see Fig. 467). Instead of (4) we then have

$$\begin{aligned} \frac{1}{k}(u_{i,j+1} - u_{ij}) = {} & \frac{1}{2h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \\ & + \frac{1}{2h^2}(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}). \end{aligned} \tag{7}$$

Multiplying by $2k$ and writing $r = k/h^2$ as before, we collect the terms corresponding to time row $j + 1$ on the left and the terms corresponding to time row j on the right:

$$(2 + 2r)u_{i,j+1} - r(u_{i+1,j+1} + u_{i-1,j+1}) = (2 - 2r)u_{ij} + r(u_{i+1,j} + u_{i-1,j}). \tag{8}$$

How do we use (8)? In general, the three values on the left are unknown, whereas the three values on the right are known. If we divide the x-interval $0 \leqq x \leqq 1$ in (1) into n equal intervals, we have $n - 1$ internal mesh points per time row (see Fig. 465, where $n = 4$). Then for $j = 0$ and $i = 1, \cdots, n - 1$, formula (8) gives a linear system of $n - 1$ equations for the $n - 1$ unknown values $u_{11}, u_{21}, \cdots, u_{n-1,1}$ in the first time row in terms of the initial values $u_{00}, u_{10}, \cdots, u_{n0}$ and the boundary values $u_{01} (= 0), u_{n1} (= 0)$. Similarly for $j = 1, j = 2$, and so on; that is, for each time row we have to solve such a linear system of $n - 1$ equations resulting from (8).

Although $r = k/h^2$ is no longer restricted, smaller r will still give better results. In practice, one chooses a k by which one can save a considerable amount of work, without

[5]JOHN CRANK (1916–2006), English mathematician and physicist at Courtaulds Fundamental Research Laboratory, professor at Brunel University, England. Student of Sir WILLIAM LAWRENCE BRAGG (1890–1971), Australian British physicist, who with his father, Sir WILLIAM HENRY BRAGG (1862–1942) won the Nobel Prize in physics in 1915 for their fundamental work in X-ray crystallography. (This is the only case where a father and a son shared the Nobel Prize for the same research. Furthermore, W. L. Bragg is the youngest Nobel laureate ever.) PHYLLIS NICOLSON (1917–1968), English mathematician, professor at the University of Leeds, England.

making r too large. For instance, often a good choice is $r = 1$ (which would be impossible in the previous method). Then (8) becomes simply

$$4u_{i,j+1} - u_{i+1,j+1} - u_{i-1,j+1} = u_{i+1,j} + u_{i-1,j}. \tag{9}$$

Time row $j + 1$

k

Time row j

h h

Fig. 467. The six points in the Crank–Nicolson formulas (7) and (8)

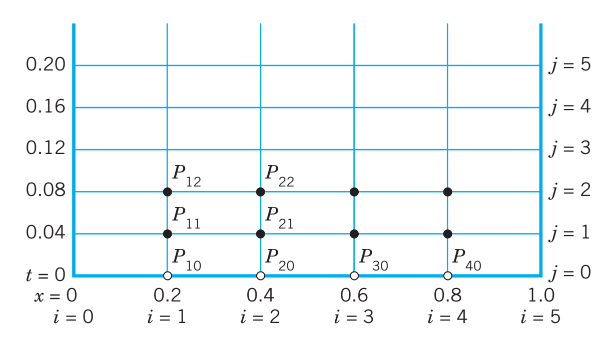


Fig. 468. Grid in Example 1

EXAMPLE 1 Temperature in a Metal Bar. Crank–Nicolson Method, Explicit Method

Consider a laterally insulated metal bar of length 1 and such that $c^2 = 1$ in the heat equation. Suppose that the ends of the bar are kept at temperature $u = 0°C$ and the temperature in the bar at some instant—call it $t = 0$—is $f(x) = \sin \pi x$. Applying the Crank–Nicolson method with $h = 0.2$ and $r = 1$, find the temperature $u(x, t)$ in the bar for $0 \leqq t \leqq 0.2$. Compare the results with the exact solution. Also apply (5) with an r satisfying (6), say, $r = 0.25$, and with values not satisfying (6), say, $r = 1$ and $r = 2.5$.

Solution by Crank–Nicolson. Since $r = 1$, formula (8) takes the form (9). Since $h = 0.2$ and $r = k/h^2 = 1$, we have $k = h^2 = 0.04$. Hence we have to do 5 steps. Figure 468 shows the grid. We shall need the initial values

$$u_{10} = \sin 0.2\pi = 0.587785, \qquad u_{20} = \sin 0.4\pi = 0.951057.$$

Also, $u_{30} = u_{20}$ and $u_{40} = u_{10}$. (Recall that u_{10} means u at P_{10} in Fig. 468, etc.) In each time row in Fig. 468 there are 4 internal mesh points. Hence in each time step we would have to solve 4 equations in 4 unknowns. But since the initial temperature distribution is symmetric with respect to $x = 0.5$, and $u = 0$ at both ends for all t, we have $u_{31} = u_{21}$, $u_{41} = u_{11}$ in the first time row and similarly for the other rows. This reduces each system to 2 equations in 2 unknowns. By (9), since $u_{31} = u_{21}$ and $u_{01} = 0$, for $j = 0$ these equations are

$$\begin{aligned} (i = 1) \qquad & 4u_{11} - u_{21} \qquad\qquad = u_{00} + u_{20} = 0.951057 \\ (i = 2) \qquad & -u_{11} + 4u_{21} - u_{21} = u_{10} + u_{20} = 1.538842. \end{aligned}$$

The solution is $u_{11} = 0.399274$, $u_{21} = 0.646039$. Similarly, for time row $j = 1$ we have the system

$$\begin{aligned} (i = 1) \qquad & 4u_{12} - u_{22} = u_{01} + u_{21} = 0.646039 \\ (i = 2) \qquad & -u_{12} + 3u_{22} = u_{11} + u_{21} = 1.045313. \end{aligned}$$

The solution is $u_{12} = 0.271221$, $u_{22} = 0.438844$, and so on. This gives the temperature distribution (Fig. 469):

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.00	0	0.588	0.951	0.951	0.588	0
0.04	0	0.399	0.646	0.646	0.399	0
0.08	0	0.271	0.439	0.439	0.271	0
0.12	0	0.184	0.298	0.298	0.184	0
0.16	0	0.125	0.202	0.202	0.125	0
0.20	0	0.085	0.138	0.138	0.085	0

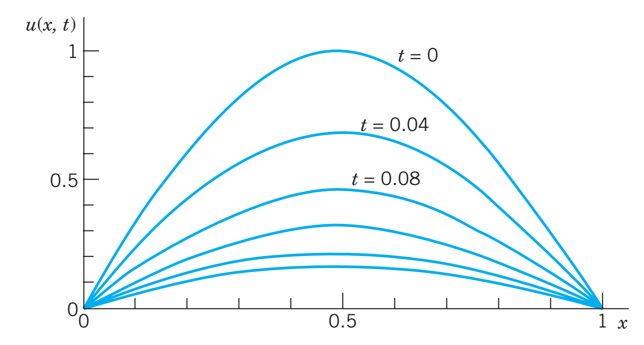


Fig. 469. Temperature distribution in the bar in Example 1

Comparison with the exact solution. The present problem can be solved exactly by separating variables (Sec. 12.5); the result is

$$u(x, t) = \sin \pi x \, e^{-\pi^2 t}. \tag{10}$$

Solution by the explicit method (5) with $r = 0.25$. For $h = 0.2$ and $r = k/h^2 = 0.25$ we have $k = rh^2 = 0.25 \cdot 0.04 = 0.01$. Hence we have to perform 4 times as many steps as with the Crank–Nicolson method! Formula (5) with $r = 0.25$ is

$$u_{i,j+1} = 0.25(u_{i-1,j} + 2u_{ij} + u_{i+1,j}). \tag{11}$$

We can again make use of the symmetry. For $j = 0$ we need $u_{00} = 0$, $u_{10} = 0.587785$ (see p. 939), $u_{20} = u_{30} = 0.951057$ and compute

$$u_{11} = 0.25(u_{00} + 2u_{10} + u_{20}) = 0.531657$$

$$u_{21} = 0.25(u_{10} + 2u_{20} + u_{30}) = 0.25(u_{10} + 3u_{20}) = 0.860239.$$

Of course we can omit the boundary terms $u_{01} = 0$, $u_{02} = 0, \cdots$ from the formulas. For $j = 1$ we compute

$$u_{12} = 0.25(2u_{11} + u_{21}) = 0.480888$$

$$u_{22} = 0.25(u_{11} + 3u_{21}) = 0.778094$$

and so on. We have to perform 20 steps instead of the 5 CN steps, but the numeric values show that the accuracy is only about the same as that of the Crank–Nicolson values CN. The exact 3D-values follow from (10).

t	$x = 0.2$			$x = 0.4$		
	CN	By (11)	Exact	CN	By (11)	Exact
0.04	0.399	0.393	0.396	0.646	0.637	0.641
0.08	0.271	0.263	0.267	0.439	0.426	0.432
0.12	0.184	0.176	0.180	0.298	0.285	0.291
0.16	0.125	0.118	0.121	0.202	0.191	0.196
0.20	0.085	0.079	0.082	0.138	0.128	0.132

Failure of (5) with r violating (6). Formula (5) with $h = 0.2$ and $r = 1$—which violates (6)—is

$$u_{i,j+1} = u_{i-1,j} - u_{ij} + u_{i+1,j}$$

and gives very poor values; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.04	0.363	0.396	0.588	0.641
0.12	0.139	0.180	0.225	0.291
0.20	0.053	0.082	0.086	0.132

Formula (5) with an even larger $r = 2.5$ (and $h = 0.2$ as before) gives completely nonsensical results; some of these are

t	$x = 0.2$	Exact	$x = 0.4$	Exact
0.1	0.0265	0.2191	0.0429	0.3545
0.3	0.0001	0.0304	0.0001	0.0492.

PROBLEM SET 21.6

1. **Nondimensional form.** Show that the heat equation $\tilde{u}_{\tilde{t}} = c^2\tilde{u}_{\tilde{x}\tilde{x}}$, $0 \leqq \tilde{x} \leqq L$, can be transformed to the "nondimensional" standard form $u_t = u_{xx}$, $0 \leqq x \leqq 1$, by setting $x = \tilde{x}/L$, $t = c^2\tilde{t}/L^2$, $u = \tilde{u}/u_0$, where u_0 is any constant temperature.
2. **Difference equation.** Derive the difference approximation (4) of the heat equation.
3. **Explicit method.** Derive (5) by solving (4) for $u_{i,j+1}$.
4. **CAS EXPERIMENT. Comparison of Methods.**

   (a) Write programs for the explicit and the Crank—Nicolson methods.

   (b) Apply the programs to the heat problem of a laterally insulated bar of length 1 with $u(x, 0) = \sin \pi x$ and $u(0, t) = u(1, t) = 0$ for all t, using $h = 0.2$, $k = 0.01$ for the explicit method (20 steps), $h = 0.2$ and (9) for the Crank–Nicolson method (5 steps). Obtain exact 6D-values from a suitable series and compare.

   (c) Graph temperature curves in (b) in two figures similar to Fig. 299 in Sec. 12.7.

   (d) Experiment with smaller h (0.1, 0.05, etc.) for both methods to find out to what extent accuracy increases under systematic changes of h and k.

EXPLICIT METHOD

5. Using (5) with $h = 1$ and $k = 0.5$, solve the heat problem (1)–(3) to find the temperature at $t = 2$ in a laterally insulated bar of length 10 ft and initial temperature $f(x) = x(1 - 0.1x)$.
6. Solve the heat problem (1)–(3) by the explicit method with $h = 0.2$ and $k = 0.01$, 8 time steps, when $f(x) = x$ if $0 \leqq x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} \leqq x \leqq 1$. Compare with the 3S-values 0.108, 0.175 for $t = 0.08$, $x = 0.2, 0.4$ obtained from the series (2 terms) in Sec. 12.5.
7. The accuracy of the explicit method depends on $r\ (\leqq \frac{1}{2})$. Illustrate this for Prob. 6, choosing $r = \frac{1}{2}$ (and $h = 0.2$ as before). Do 4 steps. Compare the values for $t = 0.04$ and 0.08 with the 3S-values in Prob. 6, which are 0.156, 0.254 ($t = 0.04$), 0.105, 0.170 ($t = 0.08$).

8. In a laterally insulated bar of length 1 let the initial temperature be $f(x) = x$ if $0 \leqq x < 0.5$, $f(x) = 1 - x$ if $0.5 \leqq x \leqq 1$. Let (1) and (3) hold. Apply the explicit method with $h = 0.2$, $k = 0.01$, 5 steps. Can you expect the solution to satisfy $u(x, t) = u(1 - x, t)$ for all t?

9. Solve Prob. 8 with $f(x) = x$ if $0 \leqq x \leqq 0.2$, $f(x) = 0.25(1 - x)$ if $0.2 < x \leqq 1$, the other data being as before.

10. Insulated end. If the left end of a laterally insulated bar extending from $x = 0$ to $x = 1$ is insulated, the boundary condition at $x = 0$ is $u_n(0, t) = u_x(0, t) = 0$. Show that, in the application of the explicit method given by (5), we can compute u_{0j+1} by the formula

$$u_{0j+1} = (1 - 2r)u_{0j} + 2ru_{1j}.$$

Apply this with $h = 0.2$ and $r = 0.25$ to determine the temperature $u(x, t)$ in a laterally insulated bar extending from $x = 0$ to 1 if $u(x, 0) = 0$, the left end is insulated and the right end is kept at temperature $g(t) = \sin \frac{50}{3}\pi t$. *Hint.* Use $0 = \partial u_{0j}/\partial x = (u_{1j} - u_{-1j})/2h$.

CRANK–NICOLSON METHOD

11. Solve Prob. 9 by (9) with $h = 0.2$, 2 steps. Compare with exact values obtained from the series in Sec. 12.5 (2 terms) with suitable coefficients.

12. Solve the heat problem (1)–(3) by Crank–Nicolson for $0 \leqq t \leqq 0.20$ with $h = 0.2$ and $k = 0.04$ when $f(x) = x$ if $0 \leqq x < \frac{1}{2}$, $f(x) = 1 - x$ if $\frac{1}{2} \leqq x \leqq 1$. Compare with the exact values for $t = 0.20$ obtained from the series (2 terms) in Sec. 12.5.

13–15

Solve (1)–(3) by Crank–Nicolson with $r = 1$ (5 steps), where:

13. $f(x) = 5x$ if $0 \leqq x < 0.25$, $f(x) = 1.25(1 - x)$ if $0.25 \leqq x \leqq 1$, $h = 0.2$

14. $f(x) = x(1 - x)$, $h = 0.1$. (Compare with Prob. 15.)

15. $f(x) = x(1 - x)$, $h = 0.2$

21.7 Method for Hyperbolic PDEs

In this section we consider the numeric solution of problems involving hyperbolic PDEs. We explain a standard method in terms of a typical setting for the prototype of a hyperbolic PDE, the **wave equation**:

(1)	$u_{tt} = u_{xx}$	$0 \leqq x \leqq 1, t \geqq 0$
(2)	$u(x, 0) = f(x)$	(Given initial displacement)
(3)	$u_t(x, 0) = g(x)$	(Given initial velocity)
(4)	$u(0, t) = u(1, t) = 0$	(Boundary conditions).

Note that an equation $u_{tt} = c^2u_{xx}$ and another x-interval can be reduced to the form (1) by a linear transformation of x and t. This is similar to Sec. 21.6, Prob. 1.

For instance, (1)–(4) is the model of a vibrating elastic string with fixed ends at $x = 0$ and $x = 1$ (see Sec. 12.2). Although an analytic solution of the problem is given in (13), Sec. 12.4, we use the problem for explaining basic ideas of the numeric approach that are also relevant for more complicated hyperbolic PDEs.

Replacing the derivatives by difference quotients as before, we obtain from (1) [see (6) in Sec. 21.4 with $y = t$]

$$\text{(5)} \qquad \frac{1}{k^2}(u_{i,j+1} - 2u_{ij} + u_{i,j-1}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j})$$

where h is the mesh size in x, and k is the mesh size in t. This difference equation relates 5 points as shown in Fig. 470a. It suggests a rectangular grid similar to the grids for

parabolic equations in the preceding section. We choose $r^* = k^2/h^2 = 1$. Then u_{ij} drops out and we have

$$u_{i,j+1} = u_{i-1,j} + u_{i+1,j} - u_{1,j-1} \tag{6}$$ (Fig. 470b).

It can be shown that for $0 < r^* \leqq 1$ the present **explicit method** is stable, so that from (6) we may expect reasonable results for initial data that have no discontinuities. (For a hyperbolic PDE the latter would propagate into the solution domain—a phenomenon that would be difficult to deal with on our present grid. For unconditionally stable **implicit methods** see [E1] in App. 1.)

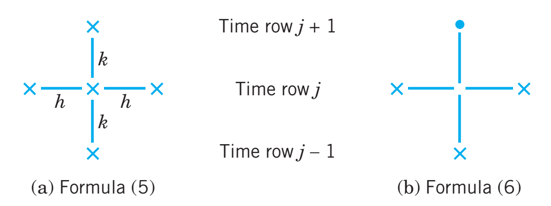


Fig. 470. Mesh points used in (5) and (6)

Equation (6) still involves 3 time steps $j-1, j, j+1$, whereas the formulas in the parabolic case involved only 2 time steps. Furthermore, we now have 2 initial conditions. So we ask how we get started and how we can use the initial condition (3). This can be done as follows.

From $u_t(x, 0) = g(x)$ we derive the difference formula

$$\frac{1}{2k}(u_{i1} - u_{i,-1}) = g_i, \qquad \text{hence} \qquad u_{i,-1} = u_{i1} - 2kg_i \tag{7}$$

where $g_i = g(ih)$. For $t = 0$, that is, $j = 0$, equation (6) is

$$u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i,-1}.$$

Into this we substitute $u_{i,-1}$ as given in (7). We obtain $u_{i1} = u_{i-1,0} + u_{i+1,0} - u_{i1} + 2kg_i$ and by simplification

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}) + kg_i, \tag{8}$$

This expresses u_{i1} in terms of the initial data. It is for the beginning only. Then use (6).

EXAMPLE 1 Vibrating String, Wave Equation

Apply the present method with $h = k = 0.2$ to the problem (1)–(4), where

$$f(x) = \sin \pi x, \qquad g(x) = 0.$$

Solution. The grid is the same as in Fig. 468, Sec. 21.6, except for the values of t, which now are $0.2, 0.4, \cdots$ (instead of $0.04, 0.08, \cdots$). The initial values $u_{00}, u_{10}, \cdots$ are the same as in Example 1, Sec. 21.6. From (8) and $g(x) = 0$ we have

$$u_{i1} = \tfrac{1}{2}(u_{i-1,0} + u_{i+1,0}).$$

From this we compute, using $u_{10} = u_{40} = \sin 0.2\pi = 0.587785$, $u_{20} = u_{30} = 0.951057$,

$$(i = 1) \quad u_{11} = \tfrac{1}{2}(u_{00} + u_{20}) = \tfrac{1}{2} \cdot 0.951057 = 0.475528$$

$$(i = 2) \quad u_{21} = \tfrac{1}{2}(u_{10} + u_{30}) = \tfrac{1}{2} \cdot 1.538842 = 0.769421$$

and $u_{31} = u_{21}$, $u_{41} = u_{11}$ by symmetry as in Sec. 21.6, Example 1. From (6) with $j = 1$ we now compute, using $u_{01} = u_{02} = \cdots = 0$,

$$(i = 1) \quad u_{12} = u_{01} + u_{21} - u_{10} = 0.769421 - 0.587785 = 0.181636$$

$$(i = 2) \quad u_{22} = u_{11} + u_{31} - u_{20} = 0.475528 + 0.769421 - 0.951057 = 0.293892,$$

and $u_{32} = u_{22}$, $u_{42} = u_{12}$ by symmetry; and so on. We thus obtain the following values of the displacement $u(x, t)$ of the string over the first half-cycle:

t	$x = 0$	$x = 0.2$	$x = 0.4$	$x = 0.6$	$x = 0.8$	$x = 1$
0.0	0	0.588	0.951	0.951	0.588	0
0.2	0	0.476	0.769	0.769	0.476	0
0.4	0	0.182	0.294	0.294	0.182	0
0.6	0	−0.182	−0.294	−0.294	−0.182	0
0.8	0	−0.476	−0.769	−0.769	−0.476	0
1.0	0	−0.588	−0.951	−0.951	−0.588	0

These values are exact to 3D (3 decimals), the exact solution of the problem being (see Sec. 12.3)

$$u(x, t) = \sin \pi x \cos \pi t.$$

The reason for the exactness follows from d'Alembert's solution (4), Sec. 12.4. (See Prob. 4, below.) ■

This is the end of Chap. 21 on numerics for ODEs and PDEs, a field that continues to develop rapidly in both applications and theoretical research. Much of the activity in the field is due to the computer serving as an invaluable tool for solving large-scale and complicated practical problems as well as for testing and experimenting with innovative ideas. These ideas could be small or major improvements on existing numeric algorithms or testing new algorithms as well as other ideas.

PROBLEM SET 21.7

VIBRATING STRING

1–3 Using the present method, solve (1)–(4) with $h = k = 0.2$ for the given initial deflection $f(x)$ and initial velocity 0 on the given t-interval.

1. $f(x) = x$ if $0 = x < \frac{1}{5}$, $f(x) = \frac{1}{4}(1 - x)$ if $\frac{1}{5} \leqq x \leqq 1$, $0 \leqq t \leqq 1$
2. $f(x) = x^2 - x^3$, $0 \leqq t \leqq 2$
3. $f(x) = 0.2(x - x^2)$, $0 \leqq t \leqq 2$

4. **Another starting formula.** Show that (12) in Sec. 12.4 gives the starting formula

$$u_{i,1} = \frac{1}{2}(u_{i+1,0} + u_{i-1,0}) + \frac{1}{2}\int_{x_i-k}^{x_i+k} g(s)\, ds$$

(where one can evaluate the integral numerically if necessary). In what case is this identical with (8)?

5. **Nonzero initial displacement and speed.** Illustrate the starting procedure when both f and g are not identically

zero, say, $f(x) = 1 - \cos 2\pi x$, $g(x) = x(1 - x)$, $h = k = 0.1$, 2 time steps.

6. Solve (1)–(3) ($h = k = 0.2$, 5 time steps) subject to $f(x) = x^2$, $g(x) = 2x$, $u_x(0, t) = 2t$, $u(1, t) = (1 + t)^2$.

7. Zero initial displacement. If the string governed by the wave equation (1) starts from its equilibrium position with initial velocity $g(x) = \sin \pi x$, what is its displacement at time $t = 0.4$ and $x = 0.2, 0.4, 0.6, 0.8$? (Use the present method with $h = 0.2$, $k = 0.2$. Use (8). Compare with the exact values obtained from (12) in Sec. 12.4.)

8. Compute approximate values in Prob. 7, using a finer grid ($h = 0.1$, $k = 0.1$), and notice the increase in accuracy.

9. Compute u in Prob. 5 for $t = 0.1$ and $x = 0.1, 0.2, \cdots, 0.9$, using the formula in Prob. 8, and compare the values.

10. Show that from d'Alembert's solution (13) in Sec.12.4 with $c = 1$ it follows that (6) in the present section gives the exact value $u_{i,j+1} = u(ih, (j + 1)h)$.

CHAPTER 21 REVIEW QUESTIONS AND PROBLEMS

1. Explain the Euler and improved Euler methods in geometrical terms. Why did we consider these methods?

2. How did we obtain numeric methods from the Taylor series?

3. What are the local and the global orders of a method? Give examples.

4. Why did we compute auxiliary values in each Runge–Kutta step? How many?

5. What is adaptive integration? How does its idea extend to Runge–Kutta?

6. What are one-step methods? Multistep methods? The underlying ideas? Give examples.

7. What does it mean that a method is not self-starting? How do we overcome this problem?

8. What is a predictor–corrector method? Give an important example.

9. What is automatic step size control? When is it needed? How is it done in practice?

10. How do we extend Runge–Kutta to systems of ODEs?

11. Why did we have to treat the main types of PDEs in separate sections? Make a list of types of problems and numeric methods.

12. When and how did we use finite differences? Give as many details as you can remember without looking into the text.

13. How did we approximate the Laplace and Poisson equations?

14. How many initial conditions did we prescribe for the wave equation? For the heat equation?

15. Can we expect a difference equation to give the exact solution of the corresponding PDE?

16. In what method for PDEs did we have convergence problems?

17. Solve $y' = y$, $y(0) = 1$ by Euler's method, 10 steps, $h = 0.1$.

18. Do Prob. 17 with $h = 0.01$, 10 steps. Compute the errors. Compare the error for $x = 0.1$ with that in Prob. 17.

19. Solve $y' = 1 + y^2$, $y(0) = 0$ by the improved Euler method, $h = 0.1$, 10 steps.

20. Solve $y' + y = (x + 1)^2$, $y(0) = 3$ by the improved Euler method, 10 steps with $h = 0.1$. Determine the errors.

21. Solve Prob. 19 by RK with $h = 0.1$, 5 steps. Compute the error. Compare with Prob. 19.

22. Fair comparison. Solve $y' = 2x^{-1}\sqrt{y - \ln x} + x^{-1}$, $y(1) = 0$ for $1 \leqq x \leqq 1.8$ **(a)** by the Euler method with $h = 0.1$, **(b)** by the improved Euler method with $h = 0.2$, and **(c)** by RK with $h = 0.4$. Verify that the exact solution is $y = (\ln x)^2 + \ln x$. Compute and compare the errors. Why is the comparison fair?

23. Apply the Adams–Moulton method to $y' = \sqrt{1 - y^2}$, $y(0) = 0$, $h = 0.2$, $x = 0, \cdots, 1$, starting with 0.198668, 0.389416, 0.564637.

24. Apply the A–M method to $y' = (x + y - 4)^2$, $y(0) = 4$, $h = 0.2$, $x = 0, \cdots, 1$, starting with 4.00271, 4.02279, 4.08413.

25. Apply Euler's method for systems to $y'' = x^2y$, $y(0) = 1$, $y'(0) = 0$, $h = 0.1$, 5 steps.

26. Apply Euler's method for systems to $y_1' = y_2$, $y_2' = -4y_1$, $y_1(0) = 2$, $y_2(0) = 0$, $h = 0.2$, 10 steps. Sketch the solution.

27. Apply Runge–Kutta for systems to $y'' + y = 2e^x$, $y(0) = 0$, $y'(0) = 1$, $h = 0.2$, 5 steps. Determine the errors.

28. Apply Runge–Kutta for systems to $y_1' = 6y_1 + 9y_2$, $y_2' = y_1 + 6y_2$, $y_1(0) = -3$, $y_2(0) = -3$, $h = 0.05$, 3 steps.

29. Find rough approximate values of the electrostatic potential at P_{11}, P_{12}, P_{13} in Fig. 471 that lie in a field between conducting plates (in Fig. 471 appearing as sides of a rectangle) kept at potentials 0 and 220 V as shown. (Use the indicated grid.)

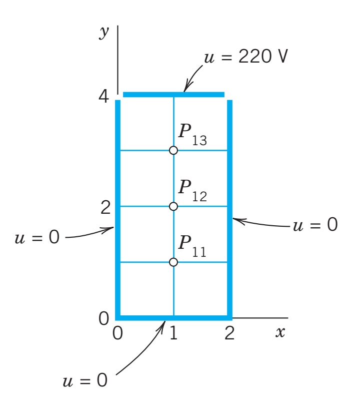


Fig. 471. Problem 29

30. A laterally insulated homogeneous bar with ends at $x = 0$ and $x = 1$ has initial temperature 0. Its left end is kept at 0, whereas the temperature at the right end varies sinusoidally according to

$$u(t, 1) = g(t) = \sin \tfrac{25}{3} \pi t.$$

Find the temperature $u(x, t)$ in the bar [solution of (1) in Sec. 21.6] by the explicit method with $h = 0.2$ and $r = 0.5$ (one period, that is, $0 \leqq t \leqq 0.24$).

31. Find the solution of the vibrating string problem $u_{tt} = u_{xx}$, $u(x, 0) = x(1 - x)$, $u_t = 0$, $u(0, t) = u(1, t) = 0$ by the method in Sec. 21.7 with $h = 0.1$ and $k = 0.1$ for $t = 0.3$.

32–34 POTENTIAL

Find the potential in Fig. 472, using the given grid and the boundary values:

32. $u(P_{01}) = u(P_{03}) = u(P_{41}) = u(P_{43}) = 200$, $u(P_{10}) = u(P_{30}) = -400$, $u(P_{20}) = 1600$, $u(P_{02}) = u(P_{42}) = u(P_{14}) = u(P_{24}) = u(P_{34}) = 0$

33. $u(P_{10}) = u(P_{30}) = 960$, $u(P_{20}) = -480$, $u = 0$ elsewhere on the boundary

34. $u = 70$ on the upper and left sides, $u = 0$ on the lower and right sides

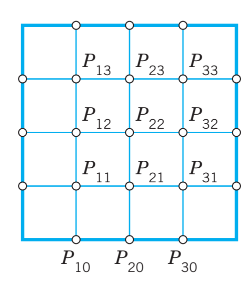


Fig. 472. Problems 32–34

35. Solve $u_t = u_{xx}$ ($0 \leqq x \leqq 1, t \geqq 0$), $u(x, 0) = x^2(1 - x)$, $u(0, t) = u(1, t) = 0$ by Crank–Nicolson with $h = 0.2$, $k = 0.04$, 5 time steps.

SUMMARY OF CHAPTER 21
Numerics for ODEs and PDEs

In this chapter we discussed numerics for ODEs (Secs. 21.1–21.3) and PDEs (Secs. 21.4–21.7). Methods for initial value problems

$$y' = f(x, y), \qquad y(x_0) = y_0 \tag{1}$$

involving a first-order ODE are obtained by truncating the Taylor series

$$y(x + h) = y(x) + hy'(x) + \frac{h^2}{2} y''(x) + \cdots$$

where, by (1), $y' = f, y'' = f' = \partial f/\partial x + (\partial f/\partial y)y'$, etc. Truncating after the term hy', we get the *Euler method,* in which we compute step by step

$$y_{n+1} = y_n + hf(x_n, y_n) \qquad (n = 0, 1, \cdots). \tag{2}$$

Taking one more term into account, we obtain the *improved Euler method.* Both methods show the basic idea but are too inaccurate in most cases.

Truncating after the term in h^4, we get the important classical **Runge–Kutta (RK) method** of fourth order. The crucial idea in this method is the replacement of the cumbersome evaluation of derivatives by the evaluation of $f(x, y)$ at suitable points (x, y); thus in each step we first compute four auxiliary quantities (Sec. 21.1)

$$\begin{aligned} k_1 &= hf(x_n, y_n) \\ k_2 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_1) \\ k_3 &= hf(x_n + \tfrac{1}{2}h, y_n + \tfrac{1}{2}k_2) \\ k_4 &= hf(x_n + h, y_n + k_3) \end{aligned} \tag{3a}$$

and then the new value

$$y_{n+1} = y_n + \tfrac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4). \tag{3b}$$

Error and step size control are possible by step halving or by **RKF** (Runge–Kutta–Fehlberg).

The methods in Sec. 21.1 are **one-step methods** since they get y_{n+1} from the result y_n of a single step. A **multistep method** (Sec. 21.2) uses the values of $y_n, y_{n-1}, \cdots$ of several steps for computing y_{n+1}. Integrating cubic interpolation polynomials gives the **Adams–Bashforth predictor** (Sec. 21.2)

$$y^*_{n+1} = y_n + \tfrac{1}{24}h(55f_n - 59f_{n-1} + 37f_{n-2} - 9f_{n-3}) \tag{4a}$$

where $f_j = f(x_j, y_j)$, and an **Adams–Moulton corrector** (the actual new value)

$$y_{n+1} = y_n + \tfrac{1}{24}h(9f^*_{n+1} + 19f_n - 5f_{n-1} + f_{n-2}), \tag{4b}$$

where $f^*_{n+1} = f(x_{n+1}, y^*_{n+1})$. Here, to get started, y_1, y_2, y_3 must be computed by the Runge–Kutta method or by some other accurate method.

Section 19.3 concerned the extension of Euler and RK methods to systems

$$\mathbf{y}' = \mathbf{f}(x, \mathbf{y}), \qquad \text{thus} \qquad y'_j = f_j(x, y_1, \cdots, y_m), \qquad j = 1, \cdots, m.$$

This includes single mth-order ODEs, which are reduced to systems. Second-order equations can also be solved by **RKN** (Runge–Kutta–Nyström) **methods**. These are particularly advantageous for $y'' = f(x, y)$ with f not containing y'.

Numeric methods for PDEs are obtained by replacing partial derivatives by difference quotients. This leads to approximating difference equations, for the **Laplace equation** to

$$u_{i+1,j} + u_{i,j+1} + u_{i-1,j} + u_{i,j-1} - 4u_{ij} = 0 \qquad (5) \quad \text{(Sec. 21.4)}$$

for the **heat equation** to

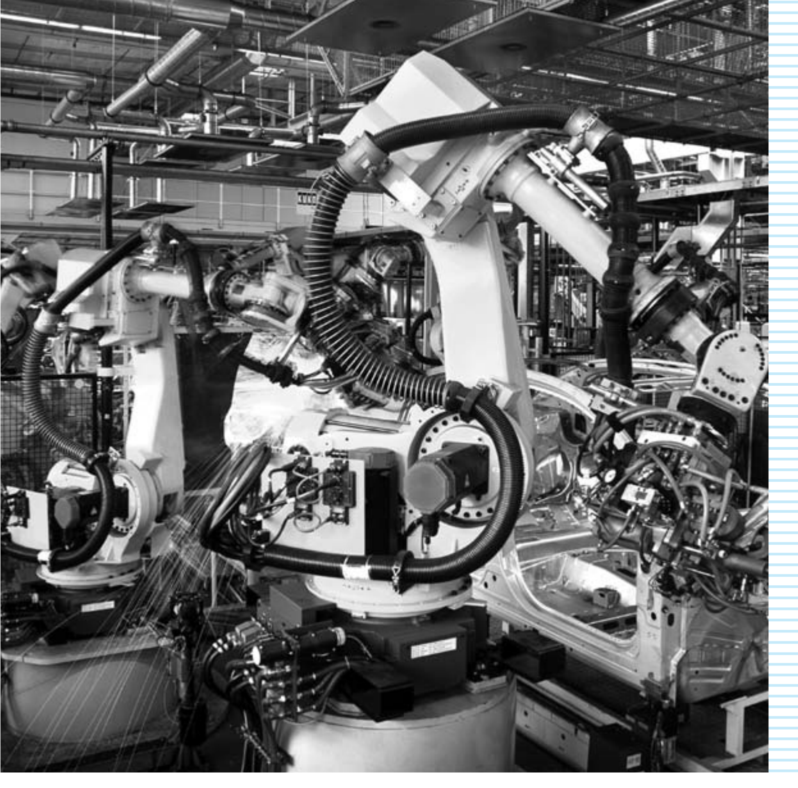
$$\frac{1}{k}(u_{i,j+1} - u_{ij}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad (6) \quad \text{(Sec. 21.6)}$$

and for the **wave equation** to

$$\frac{1}{k^2}(u_{i,j+1} - 2u_{i,j} + u_{i,j-1}) = \frac{1}{h^2}(u_{i+1,j} - 2u_{ij} + u_{i-1,j}) \qquad (7) \quad \text{(Sec. 21.7);}$$

here h and k are the mesh sizes of a grid in the x- and y-directions, respectively, where in (6) and (7) the variable y is time t.

These PDEs are *elliptic, parabolic,* and *hyperbolic,* respectively. Corresponding numeric methods differ, for the following reason. For elliptic PDEs we have boundary value problems, and we discussed for them the ***Gauss–Seidel method*** (also known as ***Liebmann's method***) and the ***ADI method*** (Secs. 21.4, 21.5). For parabolic PDEs we are given one initial condition and boundary conditions, and we discussed an *explicit method* and the ***Crank–Nicolson method*** (Sec. 21.6). For hyperbolic PDEs, the problems are similar but we are given a second initial condition (Sec. 21.7).



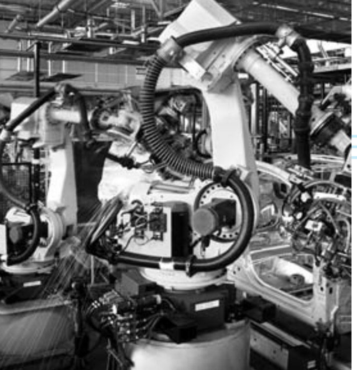
PART F

Optimization, Graphs

The material of Part F is particularly useful in modeling large-scale real-world problems. Just as it is in numerics in Part E, where the greater availability of quality software and computing power is a deciding factor in the continued growth of the field, so it is also in the fields of optimization and combinatorial optimization. Problems, such as optimizing production plans for different industries (microchips, pharmaceuticals, cars, aluminum, steel, chemicals), optimizing usage of transportation systems (usage of runways in airports, tracks of subways), efficiency in running of power plants, optimal shipping (delivery services, shipping of containers, shipping goods from factories to warehouses and from warehouses to stores), designing optimal financial portfolios, and others are all examples where the size of the problem usually requires the use of optimization software. More recently, environmental concerns have put new aspects into the picture, where an important concern, added to these problems, is the minimization of environmental impact. The main task becomes to model these problems correctly. The purpose of Part F is to introduce the main ideas and methods of unconstrained and constrained optimization (Chap. 22), and graphs and combinatorial optimization (Chap. 23).

Chapter 22 introduces unconstrained optimization by the method of ***steepest descent*** and constrained optimization by the versatile ***simplex method***. The simplex method (Secs. 22.3, 22.4) is very useful for solving many linear optimization problems (also called linear programming problems).

Graphs let us model problems in transportation logistics, efficient use of communication networks, best assignment of workers to jobs, and others. We consider shortest path problems (Secs. 22.2, 22.3), shortest spanning trees (Secs. 23.4, 23.5), flow problems in networks (Secs. 23.6, 23.7), and assignment problems (Sec. 23.8). We discuss algorithms of Moore, Dijkstra (both for shortest path), Kruskal, Prim (shortest spanning trees), and Ford–Fulkerson (for flow).



Unconstrained Optimization. Linear Programming

Optimization is a general term used to describe types of problems and solution techniques that are concerned with the best ("optimal") allocation of limited resources in projects. The problems are called optimization problems and the methods optimization methods. Typical problems are concerned with planning and making decisions, such as selecting an optimal production plan. A company has to decide how many units of each product from a choice of (distinct) products it should make. The objective of the company may be to maximize overall profit when the different products have different individual profits. In addition, the company faces certain limitations (constraints). It may have a certain number of machines, it takes a certain amount of time and usage of these machines to make a product, it requires a certain number of workers to handle the machines, and other possible criteria. To solve such a problem, you assign the first variable to number of units to be produced of the first product, the second variable to the second product, up to the number of different (distinct) products the company makes. When you multiply these, for example, by the price, you obtain a linear function called the objective function. You also express the constraints in terms of these variables, thereby obtaining several inequalities, called the constraints. Because the variables in the objective function also occur in the constraints, the objective function and the constraints are tied mathematically to each other and you have set up a linear optimization problem, also called a ***linear programming problem***.

The main focus of this chapter is to set up (Sec. 22.2) and solve (Secs. 22.3, 22.4) such linear programming problems. A famous and versatile method for doing so is the simplex method. In the ***simplex method***, the objective function and the constraints are set up in the form of an augmented matrix as in Sec. 7.3, however, the method of solving such linear constrained optimization problems is a new approach.

The beauty of the simplex method is that it allows us to scale problems up to thousands or more constraints, thereby modeling real-world situations. We can start with a small model and gradually add more and more constraints. The most difficult part is modeling the problem correctly. The actual task of solving large optimization problems is done by software implementations for the simplex method or perhaps by other optimization methods.

Besides optimal production plans, problems in optimal shipping, optimal location of warehouses and stores, easing traffic congestion, efficiency in running power plants are all examples of applications of optimization. More recent applications are in minimizing environmental damages due to pollutants, carbon dioxide emissions, and other factors. Indeed, new fields of green logistics and green manufacturing are evolving and naturally make use of optimization methods.

Prerequisite: a modest working knowledge of linear systems of equations.
References and Answers to Problems: App. 1 Part F, App. 2.

22.1 Basic Concepts. Unconstrained Optimization: Method of Steepest Descent

In an **optimization problem** the objective is to *optimize* (*maximize* or *minimize*) some function f. This function f is called the **objective function**. It is the focal point or goal of our optimization problem.

For example, an objective function f to be *maximized* may be the revenue in a production of TV sets, the rate of return of a financial portfolio, the yield per minute in a chemical process, the mileage per gallon of a certain type of car, the hourly number of customers served in a bank, the hardness of steel, or the tensile strength of a rope.

Similarly, we may want to *minimize* f if f is the cost per unit of producing certain cameras, the operating cost of some power plant, the daily loss of heat in a heating system, CO_2 emissions from a fleet of trucks for freight transport, the idling time of some lathe, or the time needed to produce a fender.

In most optimization problems the objective function f depends on several variables

$$x_1, \cdots, x_n.$$

These are called **control variables** because we can "control" them, that is, choose their values.

For example, the yield of a chemical process may depend on pressure x_1 and temperature x_2. The efficiency of a certain air-conditioning system may depend on temperature x_1, air pressure x_2, moisture content x_3, cross-sectional area of outlet x_4, and so on.

Optimization theory develops methods for optimal choices of $x_1, \cdots, x_n$, which maximize (or minimize) the objective function f, that is, methods for finding optimal values of $x_1, \cdots, x_n$.

In many problems the choice of values of $x_1, \cdots, x_n$ is not entirely free but is subject to some **constraints**, that is, additional restrictions arising from the nature of the problem and the variables.

For example, if x_1 is production cost, then $x_1 \geqq 0$, and there are many other variables (time, weight, distance traveled by a salesman, etc.) that can take nonnegative values only. Constraints can also have the form of equations (instead of inequalities).

We first consider **unconstrained optimization** in the case of a function $f(x_1, \cdots, x_n)$. We also write $\mathbf{x} = (x_1, \cdots, x_n)$ and $f(\mathbf{x})$, for convenience.

By definition, f has a **minimum** at a point $\mathbf{x} = \mathbf{X}_0$ in a region R (where f is defined) if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Similarly, f has a **maximum** at $\mathbf{X}_0$ in R if

$$f(\mathbf{x}) \leqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in R. Minima and maxima together are called **extrema**.

Furthermore, f is said to have a **local minimum** at $\mathbf{X}_0$ if

$$f(\mathbf{x}) \geqq f(\mathbf{X}_0)$$

for all $\mathbf{x}$ in a neighborhood of $\mathbf{X}_0$, say, for all $\mathbf{x}$ satisfying

$$|\mathbf{x} - \mathbf{X}_0| = [(x_1 - X_1)^2 + \cdots + (x_n - X_n)^2]^{1/2} < r,$$

where $\mathbf{X}_0 = (X_1, \cdots, X_n)$ and $r > 0$ is sufficiently small.

Similarly, f has a **local maximum** at $\mathbf{X}_0$ if $f(\mathbf{x}) \leqq f(\mathbf{X}_0)$ for all $\mathbf{x}$ satisfying $|\mathbf{x} - \mathbf{X}_0| < r$.

If f is differentiable and has an extremum at a point $\mathbf{X}_0$ in the *interior of a region R* (that is, not on the boundary), then the partial derivatives $\partial f/\partial x_1, \cdots, \partial f/\partial x_n$ must be zero at $\mathbf{X}_0$. These are the components of a vector that is called the **gradient** of f and denoted by grad f or ∇f. (For $n = 3$ this agrees with Sec. 9.7.) Thus

$$\nabla f(\mathbf{X}_0) = \mathbf{0}. \tag{1}$$

A point $\mathbf{X}_0$ at which (1) holds is called a **stationary point** of f.

Condition (1) is necessary for an extremum of f at $\mathbf{X}_0$ in the interior of R, but is not sufficient. Indeed, if $n = 1$, then for $y = f(x)$, condition (1) is $y' = f'(X_0) = 0$; and, for instance, $y = x^3$ satisfies $y' = 3x^2 = 0$ at $x = X_0 = 0$ where f has no extremum but a point of inflection. Similarly, for $f(\mathbf{x}) = x_1x_2$ we have $\nabla f(\mathbf{0}) = \mathbf{0}$, and f does not have an extremum but has a saddle point at $\mathbf{0}$. Hence, after solving (1), one must still find out whether one has obtained an extremum. In the case $n = 1$ the conditions $y'(X_0) = 0$, $y''(X_0) > 0$ guarantee a local minimum at X_0 and the conditions $y'(X_0) = 0$, $y''(X_0) < 0$ a local maximum, as is known from calculus. For $n > 1$ there exist similar criteria. However, in practice, even solving (1) will often be difficult. For this reason, one generally prefers solution by iteration, that is, by a search process that starts at some point and moves stepwise to points at which f is smaller (if a minimum of f is wanted) or larger (in the case of a maximum).

The **method of steepest descent** or **gradient method** is of this type. We present it here in its standard form. (For refinements see Ref. [E25] listed in App. 1.)

The idea of this method is to find a minimum of $f(\mathbf{x})$ by repeatedly computing minima of a function $g(t)$ of a single variable t, as follows. Suppose that f has a minimum at $\mathbf{X}_0$ and we start at a point $\mathbf{x}$. Then we look for a minimum of f closest to $\mathbf{x}$ along the straight line in the direction of $-\nabla f(\mathbf{x})$, which is the direction of steepest descent (= direction of maximum decrease) of f at $\mathbf{x}$. That is, we determine the value of t and the corresponding point

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) \tag{2}$$

at which the function

$$g(t) = f(\mathbf{z}(t)) \tag{3}$$

has a minimum. We take this $\mathbf{z}(t)$ as our next approximation to $\mathbf{X}_0$.

EXAMPLE 1 Method of Steepest Descent

Determine a minimum of

$$f(\mathbf{x}) = x_1^2 + 3x_2^2, \tag{4}$$

starting from $\mathbf{x}_0 = (6, 3) = 6\mathbf{i} + 3\mathbf{j}$ and applying the method of steepest descent.

Solution. Clearly, inspection shows that $f(\mathbf{x})$ has a minimum at $\mathbf{0}$. Knowing the solution gives us a better feel of how the method works. We obtain $\nabla f(\mathbf{x}) = 2x_1\mathbf{i} + 6x_2\mathbf{j}$ and from this

$$\mathbf{z}(t) = \mathbf{x} - t\nabla f(\mathbf{x}) = (1 - 2t)x_1\mathbf{i} + (1 - 6t)x_2\mathbf{j}$$

$$g(t) = f(\mathbf{z}(t)) = (1 - 2t)^2x_1^2 + 3(1 - 6t)^2x_2^2.$$

We now calculate the derivative

$$g'(t) = 2(1 - 2t)x_1^2(-2) + 6(1 - 6t)x_2^2(-6),$$

set $g'(t) = 0$, and solve for t, finding

$$t = \frac{x_1^2 + 9x_2^2}{2x_1^2 + 54x_2^2}.$$

Starting from $x_0 = 6\mathbf{i} + 3\mathbf{j}$, we compute the values in Table 22.1, which are shown in Fig. 473.

Figure 473 suggests that in the case of slimmer ellipses ("a long narrow valley"), convergence would be poor. You may confirm this by replacing the coefficient 3 in (4) with a large coefficient. For more sophisticated descent and other methods, some of them also applicable to vector functions of vector variables, we refer to the references listed in Part F of App. 1; see also [E25]. ■

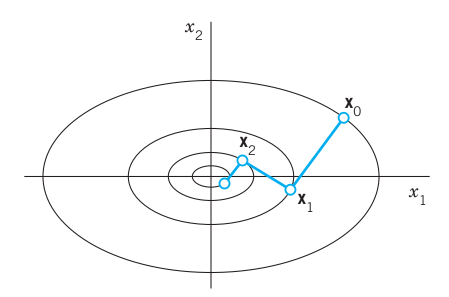


Fig. 473. Method of steepest descent in Example 1

Table 22.1 Method of Steepest Descent, Computations in Example 1

n	$\mathbf{x}$		t	$1 - 2t$	$1 - 6t$
0	6.000	3.000	0.210	0.581	−0.258
1	3.484	−0.774	0.310	0.381	−0.857
2	1.327	0.664	0.210	0.581	−0.258
3	0.771	−0.171	0.310	0.381	−0.857
4	0.294	0.147	0.210	0.581	−0.258
5	0.170	−0.038	0.310	0.381	−0.857
6	0.065	0.032			

PROBLEM SET 22.1

1. **Orthogonality.** Show that in Example 1, successive gradients are orthogonal (perpendicular). Why?

2. What happens if you apply the method of steepest descent to $f(\mathbf{x}) = x_1^2 + x_2^2$? First guess, then calculate.

3–9 STEEPEST DESCENT

Do steepest descent steps when:

3. $f(\mathbf{x}) = 2x_1^2 + x_2^2 - 4x_1 + 4x_2$, $\mathbf{x}_0 = \mathbf{0}$, 3 steps

4. $f(\mathbf{x}) = x_1^2 + 0.5x_2^2 - 5.0x_1 - 3.0x_2 + 24.95$, $x_0 = (3, 4)$, 5 steps

5. $f(\mathbf{x}) = ax_1 + bx_2$, $a \neq 0, b \neq 0$. First guess, then compute.

6. $f(\mathbf{x}) = x_1^2 - x_2^2$, $\mathbf{x}_0 = (1, 2)$, 5 steps. First guess, then compute. Sketch the path. What if $\mathbf{x}_0 = (2, 1)$?

7. $f(\mathbf{x}) = x_1^2 + cx_2^2$, $\mathbf{x}_0 = (c, 1)$. Show that 2 steps give $(c, 1)$ times a factor, $-4c^2/(c^2 - 1)^2$. What can you conclude from this about the speed of convergence?

8. $f(\mathbf{x}) = x_1^2 - x_2$, $\mathbf{x}_0 = (1, 1)$; 3 steps. Sketch your path. Predict the outcome of further steps.

9. $f(\mathbf{x}) = 0.1x_1^2 + x_2^2 - 0.02x_1$, $\mathbf{x}_0 = (3, 3)$, 5 steps

10. CAS EXPERIMENT. Steepest Descent. (a) Write a program for the method.

(b) Apply your program to $f(\mathbf{x}) = x_1^2 + 4x_2^2$, experimenting with respect to speed of convergence depending on the choice of $\mathbf{x}_0$.

(c) Apply your program to $f(\mathbf{x}) = x_1^2 + x_2^4$ and to $f(\mathbf{x}) = x_1^4 + x_2^4$, $\mathbf{x}_0 = (2, 1)$. Graph level curves and your path of descent. (Try to include graphing directly in your program.)

22.2 Linear Programming

Linear programming or **linear optimization** consists of methods for solving optimization problems ***with constraints***, that is, methods for finding a maximum (or a minimum) $\mathbf{x} = (x_1, \cdots, x_n)$ of a ***linear*** objective function

$$z = f(\mathbf{x}) = a_1x_1 + a_2x_2 + \cdots + a_nx_n$$

satisfying the constraints. The latter are **linear inequalities**, such as $3x_1 + 4x_2 \leqq 36$, or $x_1 \geqq 0$, etc. (examples below). Problems of this kind arise frequently, almost daily, for instance, in production, inventory management, bond trading, operation of power plants, routing delivery vehicles, airplane scheduling, and so on. Progress in computer technology has made it possible to solve programming problems involving hundreds or thousands or more variables. Let us explain the setting of a linear programming problem and the idea of a "geometric" solution, so that we shall see what is going on.

EXAMPLE 1 Production Plan

Energy Savers, Inc., produces heaters of types S and L. The wholesale price is \$40 per heater for S and \$88 for L. Two time constraints result from the use of two machines M_1 and M_2. On M_1 one needs 2 min for an S heater and 8 min for an L heater. On M_2 one needs 5 min for an S heater and 2 min for an L heater. Determine production figures x_1 and x_2 for S and L, respectively (number of heaters produced per hour), so that the hourly revenue

$$z = f(\mathbf{x}) = 40x_1 + 88x_2$$

is maximum.

Solution. Production figures x_1 and x_2 must be nonnegative. Hence the objective function (to be maximized) and the four constraints are

$$
\begin{array}{lll}
(0) & z = 40x_1 + 88x_2 & \\
(1) & 2x_1 + 8x_2 \leqq 60 & \text{min time on machine } M_1 \\
(2) & 5x_1 + 2x_2 \leqq 60 & \text{min time on machine } M_2 \\
(3) & x_1 \geqq 0 & \\
(4) & x_2 \geqq 0. &
\end{array}
$$

Figure 474 shows (0)–(4) as follows. Constancy lines

$$z = \text{const}$$

are marked (0). These are **lines of constant revenue**. Their slope is $-40/88 = -5/11$. To increase z we must move the line upward (parallel to itself), as the arrow shows. Equation (1) with the equality sign is marked (1). It intersects the coordinate axes at $x_1 = 60/2 = 30$ (set $x_2 = 0$) and $x_2 = 60/8 = 7.5$ (set $x_1 = 0$). The arrow marks the side on which the points (x_1, x_2) lie that satisfy the inequality in (1). Similarly for Eqs. (2)–(4). The blue quadrangle thus obtained is called the **feasibility region**. It is the set of all **feasible solutions**, meaning

solutions that satisfy all four constraints. The figure also lists the revenue at O, A, B, C. The optimal solution is obtained by moving the line of constant revenue up as much as possible without leaving the feasibility region completely. Obviously, this optimum is reached when that line passes through B, the intersection (10, 5) of (1) and (2). We see that the optimal revenue

$$z_{\max} = 40 \cdot 10 + 88 \cdot 5 = \$840$$

is obtained by producing twice as many S heaters as L heaters. ■

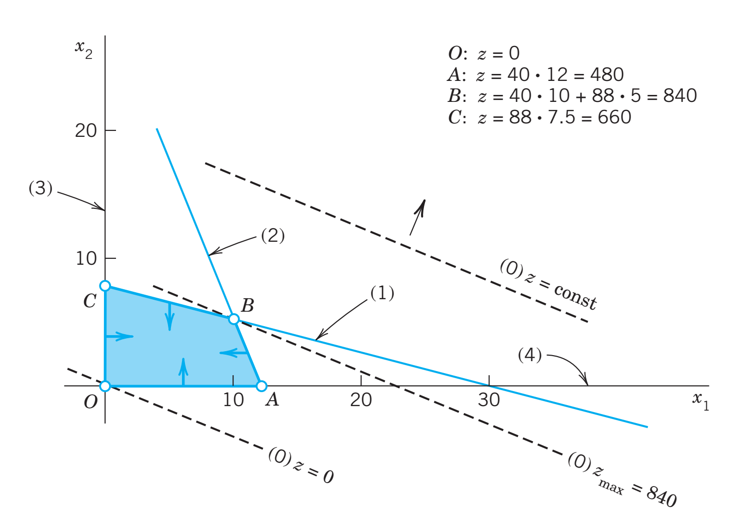


Fig. 474. Linear programming in Example 1

Note well that the problem in Example 1 or similar optimization problems *cannot* be solved by setting certain partial derivatives equal to zero, because crucial to such problems is the region in which the control variables are allowed to vary.

Furthermore, our "geometric" or graphic method illustrated in Example 1 is confined to two variables x_1, x_2. However, most practical problems involve much more than two variables, so that we need other methods of solution.

Normal Form of a Linear Programming Problem

To prepare for general solution methods, we show that constraints can be written more uniformly. Let us explain the idea in terms of (1),

$$2x_1 + 8x_2 \leqq 60.$$

This inequality implies $60 - 2x_1 - 8x_2 \geqq 0$ (and conversely), that is, the quantity

$$x_3 = 60 - 2x_1 - 8x_2$$

is nonnegative. Hence, our original inequality can now be written as an equation

$$2x_1 + 8x_2 + x_3 = 60,$$

where

$$x_3 \geqq 0.$$

x_3 is a nonnegative auxiliary variable introduced for converting inequalities to equations. Such a variable is called a **slack variable**, because it "takes up the slack" or difference between the two sides of the inequality.

EXAMPLE 2 Conversion of Inequalities by the Use of Slack Variables

With the help of two slack variables x_3, x_4 we can write the linear programming problem in Example 1 in the following form. ***Maximize***

$$f = 40x_1 + 88x_2$$

subject to the constraints

$$\begin{aligned} 2x_1 + 8x_2 + x_3 \quad\quad &= 60 \\ 5x_1 + 2x_2 \quad\quad + x_4 &= 60 \\ x_i \geqq 0 \qquad (i = 1, \cdots, 4). \end{aligned}$$

We now have $n = 4$ variables and $m = 2$ (linearly independent) equations, so that two of the four variables, for example, x_1, x_2, determine the others. Also note that each of the four sides of the quadrangle in Fig. 474 now has an equation of the form $x_i = 0$:

$$OA\colon x_2 = 0,$$

$$AB\colon x_4 = 0,$$

$$BC\colon x_3 = 0,$$

$$CO\colon x_1 = 0,$$

A vertex of the quadrangle is the intersection of two sides. Hence at a vertex, $n - m = 4 - 2 = 2$ of the variables are zero and the others are nonnegative. Thus at A we have $x_2 = 0, x_4 = 0$, and so on. ∎

Our example suggests that a general linear optimization problem can be brought to the following **normal form**. *Maximize*

$$f = c_1x_1 + c_2x_2 + \cdots + c_nx_n \tag{5}$$

subject to the constraints

$$\begin{aligned} a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\ a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\ \cdots\cdots\cdots\cdots\cdots &\cdots \\ a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\ x_i \geqq 0 \qquad (i = 1, \cdots, n) \end{aligned} \tag{6}$$

with all b_j nonnegative. (If a $b_j < 0$, multiply the equation by -1.) Here $x_1, \cdots, x_n$ include the slack variables (for which the c_j's in f are zero). We assume that the equations in (6) are linearly independent. Then, if we choose values for $n - m$ of the variables, the system uniquely determines the others. Of course, since we must have

$$x_1 \geqq 0, \cdots, x_n \geqq 0,$$

this choice is not entirely free.

Our problem also includes the **minimization** of an objective function f since this corresponds to maximizing $-f$ and thus needs no separate consideration.

An n-tuple $(x_1, \cdots, x_n)$ that satisfies all the constraints in (6) is called a *feasible point* or **feasible solution**. A feasible solution is called an **optimal solution** if, for it, the objective function f becomes maximum, compared with the values of f at all feasible solutions.

Finally, by a **basic feasible solution** we mean a feasible solution for which at least $n - m$ of the variables $x_1, \cdots, x_n$ are zero. For instance, in Example 2 we have $n = 4$, $m = 2$, and the basic feasible solutions are the four vertices O, A, B, C in Fig. 474. Here B is an optimal solution (the only one in this example).

The following theorem is fundamental.

THEOREM 1

Optimal Solution

Some optimal solution of a linear programming problem (5), (6) *is also a basic feasible solution of* (5), (6).

For a proof, see Ref. [F5], Chap. 3 (listed in App. 1). A problem can have many optimal solutions and not all of them may be *basic* feasible solutions; but the theorem guarantees that we can find an optimal solution by searching through the basic feasible solutions only. This is a great simplification; but since there are $\binom{n}{n-m} = \binom{n}{m}$ different ways of equating $n - m$ of the n variables to zero, considering all these possibilities, dropping those which are not feasible and then searching through the rest would still involve very much work, even when n and m are relatively small. Hence a systematic search is needed. We shall explain an important method of this type in the next section.

PROBLEM SET 22.2

1–6 REGIONS, CONSTRAINTS

Describe and graph the regions in the first quadrant of the x_1x_2-plane determined by the given inequalities.

1. $x_1 - 3x_2 \geqq -6$
$x_1 + x_2 \leqq 6$

2. $2x_1 - x_2 \geqq 6$
$8x_1 + 10x_2 \leqq 80$
$x_1 - 2x_2 \geqq -3$

3. $-0.5x_1 + x_2 \leqq 2$
$x_1 + x_2 \geqq 2$
$-x_1 + 5x_2 \geqq 5$

4. $-x_1 + x_2 \leqq 5$
$2x_1 + x_2 \geqq 10$
$x_2 \geqq 4$
$10x_1 + 15x_2 \leqq 150$

5. $-x_1 + x_2 \geqq 0$
$x_1 + x_2 \leqq 5$
$-2x_1 + x_2 \leqq 16$

6. $x_1 + x_2 \geqq 2$
$3x_1 + 5x_2 \geqq 15$
$2x_1 - x_2 \geqq -2$
$-x_1 + 2x_2 \leqq 10$

7. Location of maximum. Could we find a profit $f(x_1, x_2) = a_1x_1 + a_2x_2$ whose maximum is at an interior point of the quadrangle in Fig. 474? Give reason for your answer.

8. Slack variables. Why are slack variables always nonnegative? How many of them do we need?

9. What is the meaning of the slack variables x_3, x_4 in Example 2 in terms of the problem in Example 1?

10. Uniqueness. Can we always expect a unique solution (as in Example 1)?

11–16 MAXIMIZATION, MINIMIZATION

Maximize or minimize the given objective function f subject to the given constraints.

11. Maximize $f = 30x_1 + 10x_2$ in the region in Prob. 5.

12. Minimize $f = 45.0x_1 + 22.5x_2$ in the region in Prob. 4.

13. Maximize $f = 5x_1 + 25x_2$ in the region in Prob. 5.

14. Minimize $f = 5x_1 + 25x_2$ in the region in Prob. 3.

15. Maximize $f = 20x_1 + 30x_2$ subject to $4x_1 + 3x_2 \geqq 12$, $x_1 - x_2 \geqq -3$, $x_2 \leqq 6$, $2x_1 - 3x_2 \leqq 0$.

16. Maximize $f = -10x_1 + 2x_2$ subject to $x_1 \geqq 0$, $x_2 \geqq 0$, $-x_1 + x_2 \geqq -1$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 5$.

17. Maximum profit. United Metal, Inc., produces alloys B_1 (special brass) and B_2 (yellow tombac). B_1 contains 50% copper and 50% zinc. (Ordinary brass contains about 65% copper and 35% zinc.) B_2 contains 75% copper and 25% zinc. Net profits are \$120 per ton of B_1 and \$100 per ton of B_2. The daily copper supply is 45 tons. The daily zinc supply is 30 tons. Maximize the net profit of the daily production.

18. Maximum profit. The DC Drug Company produces two types of liquid pain killer, N (normal) and S (Super). Each bottle of N requires 2 units of drug A, 1 unit of drug B, and 1 unit of drug C. Each bottle of S requires 1 unit of A, 1 unit of B, and 3 units of C. The company is able to produce, each week, only 1400 units of A, 800 units of B, and 1800 units of C. The profit per bottle of N and S is \$11 and \$15, respectively. Maximize the total profit.

19. Maximum output. Giant Ladders, Inc., wants to maximize its daily total output of large step ladders by producing x_1 of them by a process P_1 and x_2 by a process P_2, where P_1 requires 2 hours of labor and 4 machine hours per ladder, and P_2 requires 3 hours of labor and 2 machine hours. For this kind of work, 1200 hours of labor and 1600 hours on the machines are, at most, available per day. Find the optimal x_1 and x_2.

20. Minimum cost. Hardbrick, Inc., has two kilns. Kiln I can produce 3000 gray bricks, 2000 red bricks, and 300 glazed bricks daily. For Kiln II the corresponding figures are 2000, 5000, and 1500. Daily operating costs of Kilns I and II are \$400 and \$600, respectively. Find the number of days of operation of each kiln so that the operation cost in filling an order of 18,000 gray, 34,000 red, and 9000 glazed bricks is minimized.

21. Maximum profit. Universal Electric, Inc., manufactures and sells two models of lamps, L_1 and L_2, the profit being \$150 and \$100, respectively. The process involves two workers W_1 and W_2 who are available for this kind of work 100 and 80 hours per month, respectively. W_1 assembles L_1 in 20 min and L_2 in 30 min. W_2 paints L_1 in 20 min and L_2 in 10 min. Assuming that all lamps made can be sold without difficulty, determine production figures that maximize the profit.

22. Nutrition. Foods A and B have 600 and 500 calories, contain 15 g and 30 g of protein, and cost \$1.80 and \$2.10 per unit, respectively. Find the minimum cost diet of at least 3900 calories containing at least 150 g of protein.

22.3 Simplex Method

From the last section we recall the following. A linear optimization problem (linear programming problem) can be written in normal form; that is:

(1) ***Maximize***

$$z = f(x) = c_1x_1 + \cdots + c_nx_n$$

subject to the constraints

(2)
$$\begin{aligned}
a_{11}x_1 + \cdots + a_{1n}x_n &= b_1 \\
a_{21}x_1 + \cdots + a_{2n}x_n &= b_2 \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\
a_{m1}x_1 + \cdots + a_{mn}x_n &= b_m \\
x_i \geqq 0 \qquad (i = 1, \cdots, n).
\end{aligned}$$

For finding an optimal solution of this problem, we need to consider only the **basic feasible solutions** (defined in Sec. 22.2), but there are still so many that we have to follow a systematic search procedure. In 1948 G. B. Dantzig[1] published an iterative method, called the **simplex method**, for that purpose. In this method, one proceeds stepwise from one basic feasible solution to another in such a way that the objective function f always increases its value. Let us explain this method in terms of the example in the last section.

In its original form the problem concerned the maximization of the objective function

$$z = 40x_1 + 88x_2$$

subject to

$$\begin{aligned} 2x_1 + 8x_2 &\leqq 60 \\ 5x_1 + 2x_2 &\leqq 60 \\ x_1 \qquad\quad &\geqq 0 \\ x_2 &\geqq 0. \end{aligned}$$

Converting the first two inequalities to equations by introducing two slack variables x_3, x_4, we obtained the **normal form** of the problem in Example 2. Together with the objective function (written as an equation $z - 40x_1 - 88x_2 = 0$) this normal form is

$$\begin{aligned} z - 40x_1 - 88x_2 \qquad\qquad &= 0 \\ 2x_1 + \ 8x_2 + x_3 \qquad &= 60 \\ 5x_1 + \ 2x_2 \qquad + x_4 &= 60 \end{aligned} \tag{3}$$

where $x_1 \geqq 0, \cdots, x_4 \geqq 0$. This is a linear system of equations. To find an optimal solution of it, we may consider its **augmented matrix** (see Sec. 7.3)

$$\mathbf{T}_0 = \begin{array}{c} \begin{matrix} z & x_1 & x_2 & x_3 & x_4 & b \end{matrix} \\ \left[\begin{array}{c|cc|cc|c} 1 & -40 & -88 & 0 & 0 & 0 \\ \hline 0 & 2 & 8 & 1 & 0 & 60 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] \end{array} \tag{4}$$

[1] GEORGE BERNARD DANTZIG (1914–2005), American mathematician, who is one of the pioneers of linear programming and inventor of the simplex method. According to Dantzig himself (see G. B. Dantzig, Linear programming: The story of how it began, in J. K. Lenestra et al., *History of Mathematical Programming: A Collection of Personal Reminiscences*. Amsterdam: Elsevier, 1991, pp. 19–31), he was particularly fascinated by Wassily Leontief's input–output model (Sec. 8.2) and invented his famous method to solve large-scale planning (logistics) problems. Besides Leontief, Dantzig credits others for their pioneering work in linear programming, that is, JOHN VON NEUMANN (1903–1957), Hungarian American mathematician, Institute for Advanced Studies, Princeton University, who made major contributions to game theory, computer science, functional analysis, set theory, quantum mechanics, ergodic theory, and other areas, the Nobel laureates LEONID VITALIYEVICH KANTOROVICH (1912–1986), Russian economist, and TJALLING CHARLES KOOPMANS (1910–1985), Dutch–American economist, who shared the 1975 Nobel Prize in Economics for their contributions to the theory of optimal allocation of resources. Dantzig was a driving force in establishing the field of linear programming and became professor of transportation sciences, operations research, and computer science at Stanford University. For his work see R. W. Cottle (ed.), *The Basic George B. Dantzig.* Palo Alto, CA: Stanford University Press, 2003.

This matrix is called a **simplex tableau** or **simplex table** (the ***initial*** *simplex table*). These are standard names. The dashed lines and the letters

$$z, \quad x_1, \quad \cdots, \quad b$$

are for ease in further manipulation.

Every simplex table contains two kinds of variables x_j. By **basic variables** we mean those whose columns have only one nonzero entry. Thus x_3, x_4 in (4) are basic variables and x_1, x_2 are **nonbasic variables**.

Every simplex table gives a basic feasible solution. It is obtained by setting the nonbasic variables to zero. Thus (4) gives the basic feasible solution

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 60/1 = 60, \qquad x_4 = 60/1 = 60, \qquad z = 0$$

with x_3 obtained from the second row and x_4 from the third.

The optimal solution (its location and value) is now obtained stepwise by pivoting, designed to take us to basic feasible solutions with higher and higher values of z until the maximum of z is reached. Here, the choice of the **pivot equation** and **pivot** are quite different from that in the Gauss elimination. The reason is that x_1, x_2, x_3, x_4 are restricted to nonnegative values.

***Step 1. Operation O_1*: Selection of the *Column* of the Pivot**

Select as the column of the pivot the first column with a negative entry in Row 1. In (4) this is Column 2 (because of the -40).

Operation O_2*: Selection of the *Row* of the Pivot.** Divide the right sides [60 and 60 in (4)] by the corresponding entries of the column just selected ($60/2 = 30$, $60/5 = 12$). Take as the pivot equation the equation that gives the ***smallest quotient. Thus the pivot is 5 because $60/5$ is smallest.

***Operation O_3*: Elimination by Row Operations.** This gives zeros above and below the pivot (as in Gauss–Jordan, Sec. 7.8).

With the notation for row operations as introduced in Sec. 7.3, the calculations in Step 1 give from the simplex table $\mathbf{T}_0$ in (4) the following simplex table (augmented matrix), with the blue letters referring to the ***previous table***.

$$\begin{array}{c} \\ (5) \qquad \mathbf{T}_1 = \end{array} \begin{array}{c} \begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\ \left[\begin{array}{c|cc|cc|c} 1 & 0 & -72 & 0 & 8 & 480 \\ \hline 0 & 0 & 7.2 & 1 & -0.4 & 36 \\ 0 & 5 & 2 & 0 & 1 & 60 \end{array}\right] \end{array} \begin{array}{l} \text{Row 1} + 8 \text{ Row 3} \\ \text{Row 2} - 0.4 \text{ Row 3} \\ \\ \end{array}$$

We see that basic variables are now x_1, x_3 and nonbasic variables are x_2, x_4. Setting the latter to zero, we obtain the basic feasible solution given by $\mathbf{T}_1$,

$$x_1 = 60/5 = 12, \qquad x_2 = 0, \qquad x_3 = 36/1 = 36, \qquad x_4 = 0, \qquad z = 480.$$

This is A in Fig. 474 (Sec. 22.2). We thus have moved from O: (0, 0) with $z = 0$ to A: (12, 0) with the greater $z = 480$. The reason for this increase is our elimination of a

term $(-40x_1)$ with a negative coefficient. Hence ***elimination is applied only to negative entries*** in Row 1 but to no others. This motivates the selection of the *column* of the pivot.

We now motivate the selection of the *row* of the pivot. Had we taken the second row of $\mathbf{T}_0$ instead (thus 2 as the pivot), we would have obtained $z = 1200$ (verify!), but this line of constant revenue $z = 1200$ lies entirely outside the feasibility region in Fig. 474. This motivates our cautious choice of the entry 5 as our pivot because it gave the smallest quotient $(60/5 = 12)$.

Step 2. The basic feasible solution given by (5) is not yet optimal because of the negative entry -72 in Row 1. Accordingly, we perform the operations O_1 to O_3 again, choosing a pivot in the column of -72.

Operation O_1. Select Column 3 of $\mathbf{T}_1$ in (5) as the column of the pivot (because $-72 < 0$).

Operation O_2. We have $36/7.2 = 5$ and $60/2 = 30$. Select 7.2 as the pivot (because $5 < 30$).

Operation O_3. Elimination by row operations gives

$$
(6) \qquad \mathbf{T}_2 = \begin{array}{c}
\begin{array}{cccccc} z & x_1 & x_2 & x_3 & x_4 & b \end{array} \\
\left[\begin{array}{c|cc|cc|c}
1 & 0 & 0 & 10 & 4 & 840 \\
\hline
0 & 0 & 7.2 & 1 & -0.4 & 36 \\
0 & 5 & 0 & -\dfrac{1}{3.6} & \dfrac{1}{0.9} & 50
\end{array}\right]
\end{array}
\begin{array}{l}
\\
\text{Row 1} + 10 \text{ Row 2} \\
\\
\text{Row 3} - \dfrac{2}{7.2}\text{ Row 2}
\end{array}
$$

We see that now x_1, x_2 are basic and x_3, x_4 nonbasic. Setting the latter to zero, we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 50/5 = 10, \qquad x_2 = 36/7.2 = 5, \qquad x_3 = 0, \qquad x_4 = 0, \qquad z = 840.$$

This is B in Fig. 474 (Sec. 22.2). In this step, z has increased from 480 to 840, due to the elimination of -72 in $\mathbf{T}_1$. Since $\mathbf{T}_2$ contains no more negative entries in Row 1, we conclude that $z = f(10, 5) = 40 \cdot 10 + 88 \cdot 5 = 840$ is the maximum possible revenue. It is obtained if we produce twice as many S heaters as L heaters. This is the solution of our problem by the simplex method of linear programming. ■

Minimization. If we want to ***minimize*** $z = f(\mathbf{x})$ (instead of maximize), we take as the columns of the pivots those whose entry in Row 1 is ***positive*** (instead of negative). In such a Column k we consider only positive entries t_{jk} and take as pivot a t_{jk} for which b_j/t_{jk} is smallest (as before). For examples, see the problem set.

PROBLEM SET 22.3

1. Verify the calculations in Example 1 of the text.

2–14 SIMPLEX METHOD

Write in normal form and solve by the simplex method, assuming all x_j to be nonnegative.

2. The problem in the example in the text with the constraints interchanged.

3. Maximize $f = 3x_1 + 2x_2$ subject to $3x_1 + 4x_2 \leqq 60$, $4x_1 + 3x_2 \leqq 60$, $10x_1 + 2x_2 \leqq 120$.

4. Maximize the daily output in producing x_1 chairs by Process P_1 and x_2 chairs by Process P_2 subject to $3x_1 + 4x_2 \leqq 550$ (machine hours), $5x_1 + 4x_2 \leqq 650$ (labor).

5. Minimize $f = 5x_1 - 20x_2$ subject to $-2x_1 + 10x_2 \leqq 5$, $2x_1 + 5x_2 \leqq 10$.

6. Prob. 19 in Sec. 22.2.

7. Suppose we produce x_1 AA batteries by Process P_1 and x_2 by Process P_2, furthermore x_3 A batteries by Process P_3 and x_4 by Process P_4. Let the profit for 100 batteries be \$10 for AA and \$20 for A. Maximize the total profit subject to the constraints

$$12x_1 + 8x_2 + 6x_3 + 4x_4 \leqq 120 \quad \text{(Material)}$$
$$3x_1 + 6x_2 + 12x_3 + 24x_4 \leqq 180 \quad \text{(Labor).}$$

8. Maximize the daily profit in producing x_1 metal frames F_1 (profit \$90 per frame) and x_2 frames F_2 (profit \$50 per frame) subject to $x_1 + 3x_2 \leqq 18$ (material), $x_1 + x_2 \leqq 10$ (machine hours), $3x_1 + x_2 \leqq 24$ (labor).

9. Maximize $f = 2x_1 + x_2 + 3x_3$ subject to $4x_1 + 3x_2 + 6x_3 = 12$.

10. Minimize $f = 4x_1 - 10x_2 - 20x_3$ subject to $3x_1 + 4x_2 + 5x_3 \leqq 60$, $2x_1 + x_2 \leqq 20$, $2x_1 + 3x_3 \leqq 30$.

11. Prob. 22 in Problem Set 22.2.

12. Maximize $f = 2x_1 + 3x_2 + x_3$ subject to $x_1 + x_2 + x_3 \leqq 4.8$, $10x_1 + x_3 \leqq 9.9$, $x_2 - x_3 \leqq 0.2$.

13. Maximize $f = 34x_1 + 29x_2 + 32x_3$ subject to $8x_1 + 2x_2 + x_3 \leqq 54$, $3x_1 + 8x_2 + 2x_3 \leqq 59$, $x_1 + x_2 + 5x_3 \leqq 39$.

14. Maximize $f = 2x_1 + 3x_2$ subject to $5x_1 + 3x_2 \leqq 105$, $3x_1 + 6x_2 \leqq 126$.

15. CAS PROJECT. Simple Method. (a) Write a program for graphing a region R in the first quadrant of the x_1x_2-plane determined by linear constraints.

(b) Write a program for maximizing $z = a_1x_1 + a_2x_2$ in R.

(c) Write a program for maximizing $z = a_1x_1 + \cdots + a_nx_n$ subject to linear constraints.

(d) Apply your programs to problems in this problem set and the previous one.

22.4 Simplex Method: Difficulties

In solving a linear optimization problem by the simplex method, we proceed stepwise from one basic feasible solution to another. By so doing, we increase the value of the objective function f. We continue this stepwise procedure, until we reach an optimal solution. This was all explained in Sec. 22.3. However, the method does not always proceed so smoothly. Occasionally, but rather infrequently in practice, we encounter two kinds of difficulties. The first one is the degeneracy and the second one concerns difficulties in starting.

Degeneracy

A **degenerate feasible solution** is a feasible solution at which more than the usual number $n - m$ of variables are zero. Here n is the number of variables (slack and others) and m the number of constraints (not counting the $x_j \geqq 0$ conditions). In the last section, $n = 4$ and $m = 2$, and the occurring basic feasible solutions were nondegenerate; $n - m = 2$ variables were zero in each such solution.

In the case of a degenerate feasible solution we do an extra elimination step in which a basic variable that is zero for that solution becomes nonbasic (and a nonbasic variable becomes basic instead). We explain this in a typical case. For more complicated cases and techniques (rarely needed in practice) see Ref. [F5] in App. 1.

EXAMPLE 1 Simplex Method, Degenerate Feasible Solution

AB Steel, Inc., produces two kinds of iron I_1, I_2 by using three kinds of raw material R_1, R_2, R_3 (scrap iron and two kinds of ore) as shown. Maximize the daily profit.

Raw Material	Raw Material Needed per Ton		Raw Material Available per Day (tons)
	Iron I_1	Iron I_2	
R_1	2	1	16
R_2	1	1	8
R_3	0	1	3.5
Net profit per ton	\$150	\$300	

Solution. Let x_1 and x_2 denote the amount (in tons) of iron I_1 and I_2, respectively, produced per day. Then our problem is as follows. Maximize

$$z = f(x) = 150x_1 + 300x_2 \tag{1}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and

$$\begin{aligned} 2x_1 + x_2 &\leqq 16 \qquad \text{(raw material } R_1\text{)} \\ x_1 + x_2 &\leqq \ \ 8 \qquad \text{(raw material } R_2\text{)} \\ x_2 &\leqq \ \ 3.5 \quad \text{(raw material } R_3\text{).} \end{aligned}$$

By introducing slack variables x_3, x_4, x_5 we obtain the normal form of the constraints

$$\begin{aligned} 2x_1 + x_2 + x_3 \qquad\qquad &= 16 \\ x_1 + x_2 \qquad + x_4 \qquad &= \ \ 8 \\ x_2 \qquad\qquad + x_5 &= \ \ 3.5 \\ x_i \geqq 0 \qquad (i = 1, \cdots, 5)&. \end{aligned} \tag{2}$$

As in the last section we obtain from (1) and (2) the initial simplex table

$$\mathbf{T}_0 = \left[\begin{array}{c|cc|ccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & -150 & -300 & 0 & 0 & 0 & 0 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 1 & 1 & 0 & 1 & 0 & 8 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array}\right]. \tag{3}$$

We see that x_1, x_2 are nonbasic variables and x_3, x_4, x_5 are basic. With $x_1 = x_2 = 0$ we have from (3) the basic feasible solution

$$x_1 = 0, \quad x_2 = 0, \quad x_3 = 16/1 = 16, \quad x_4 = 8/1 = 8, \quad x_5 = 3.5/1 = 3.5, \quad z = 0.$$

This is O: (0, 0) in Fig. 475. We have $n = 5$ variables x_j, $m = 3$ constraints, and $n - m = 2$ variables equal to zero in our solution, which thus is nondegenerate.

Step 1 of Pivoting

Operation O_1: Column Selection of Pivot. Column 2 (since $-150 < 0$).

Operation O_2: Row Selection of Pivot. $16/2 = 8$, $8/1 = 8$; $3.5/0$ is not possible. Hence we could choose Row 2 or Row 3. We choose Row 2. The pivot is 2.

***Operation O_3*:** Elimination by Row Operations. This gives the simplex table

$$(4) \qquad \mathbf{T}_1 = \begin{array}{c|cc|ccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & 0 & -225 & 75 & 0 & 0 & 1200 \\ \hline 0 & 2 & 1 & 1 & 0 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 3.5 \end{array} \qquad \begin{array}{l} \text{Row 1} + 75 \text{ Row 2} \\ \\ \text{Row 3} - \frac{1}{2} \text{ Row 2} \\ \text{Row 4} \end{array}$$

We see that the basic variables are x_1, x_4, x_5 and the nonbasic are x_2, x_3. Setting the nonbasic variables to zero, we obtain from $\mathbf{T}_1$ the basic feasible solution

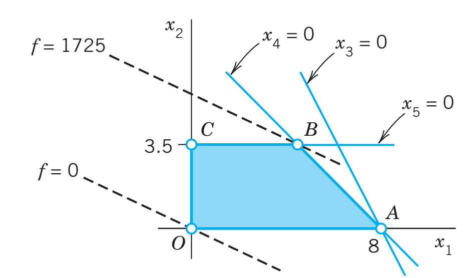


Fig. 475. Example 1, where A is degenerate

$$x_1 = 16/2 = 8, \qquad x_2 = 0, \qquad x_3 = 0, \qquad x_4 = 0/1 = 0, \qquad x_5 = 3.5/1 = 3.5, \qquad z = 1200.$$

This is A: (8, 0) in Fig. 475. This solution in degenerate because $x_4 = 0$ (in addition to $x_2 = 0, x_3 = 0$); geometrically: the straight line $x_4 = 0$ also passes through A. This requires the next step, in which x_4 will become nonbasic.

Step 2 of Pivoting

***Operation O_1*:** Column Selection of Pivot. Column 3 (since $-225 < 0$).

***Operation O_2*:** Row Selection of Pivot. $16/1 = 16, 0/\frac{1}{2} = 0$. Hence $\frac{1}{2}$ must serve as the pivot.

***Operation O_3*:** Elimination by Row Operations. This gives the following simplex table.

$$(5) \qquad \mathbf{T}_2 = \begin{array}{c|cc|ccc|c} z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & 0 & 0 & -150 & 450 & 0 & 1200 \\ \hline 0 & 2 & 0 & 2 & -2 & 0 & 16 \\ 0 & 0 & \frac{1}{2} & -\frac{1}{2} & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array} \qquad \begin{array}{l} \text{Row 1} + 450 \text{ Row 3} \\ \text{Row 2} - 2 \text{ Row 3} \\ \\ \text{Row 4} - 2 \text{ Row 3} \end{array}$$

We see that the basic variables are x_1, x_2, x_5 and the nonbasic are x_3, x_4. Hence x_4 has become nonbasic, as intended. By equating the nonbasic variables to zero we obtain from $\mathbf{T}_2$ the basic feasible solution

$$x_1 = 16/2 = 8, \qquad x_2 = 0/\tfrac{1}{2} = 0, \qquad x_3 = 0, \qquad x_4 = 0, \qquad x_5 = 3.5/1 = 3.5, \qquad z = 1200.$$

This is still A: (8, 0) in Fig. 475 and z has not increased. But this opens the way to the maximum, which we reach in the next step.

Step 3 of Pivoting

Operation O_1: Column Selection of Pivot. Column 4 (since $-150 < 0$).

Operation O_2: Row Selection of Pivot. $16/2 = 8$, $0/(-\frac{1}{2}) = 0$, $3.5/1 = 3.5$. We can take 1 as the pivot. (With $-\frac{1}{2}$ as the pivot we would not leave A. Try it.)

Operation O_3: Elimination by Row Operations. This gives the simplex table

$$\mathbf{T}_3 = \left[\begin{array}{c|cc|cccc|c} & z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\ \hline 1 & 0 & 0 & 0 & 150 & 150 & 1725 \\ \hline 0 & 2 & 0 & 0 & 2 & -2 & 9 \\ 0 & 0 & \frac{1}{2} & 0 & 0 & \frac{1}{2} & 1.75 \\ 0 & 0 & 0 & 1 & -2 & 1 & 3.5 \end{array}\right] \tag{6}$$

z	x_1	x_2	x_3	x_4	x_5	b	
1	0	0	0	150	150	1725	Row 1 + 150 Row 4
0	2	0	0	2	−2	9	Row 2 − 2 Row 4
0	0	$\frac{1}{2}$	0	0	$\frac{1}{2}$	1.75	Row 3 + $\frac{1}{2}$ Row 4
0	0	0	1	−2	1	3.5	

We see that basic variables are x_1, x_2, x_3 and nonbasic x_4, x_5. Equating the latter to zero we obtain from $\mathbf{T}_3$ the basic feasible solution

$$x_1 = 9/2 = 4.5, \quad x_2 = 1.75/\tfrac{1}{2} = 3.5, \quad x_3 = 3.5/1 = 3.5, \quad x_4 = 0, \quad x_5 = 0, \quad z = 1725.$$

This is B: (4.5, 3.5) in Fig. 475. Since Row 1 of $\mathbf{T}_3$ has no negative entries, we have reached the maximum daily profit $z_{max} = f(4.5, 3.5) = 150 \cdot 4.5 + 300 \cdot 3.5 = \$1725$. This is obtained by using 4.5 tons of iron I_1 and 3.5 tons of iron I_2. ■

Difficulties in Starting

As a second kind of difficulty, it may sometimes be hard to find a basic feasible solution to start from. In such a case the idea of an **artificial variable** (or several such variables) is helpful. We explain this method in terms of a typical example.

EXAMPLE 2 Simplex Method: Difficult Start, Artificial Variable

Maximize

$$z = f(\mathbf{x}) = 2x_1 + x_2 \tag{7}$$

subject to the constraints $x_1 \geqq 0$, $x_2 \geqq 0$ and (Fig. 476)

$$\begin{aligned} x_1 - \tfrac{1}{2}x_2 &\geqq 1 \\ x_1 - x_2 &\leqq 2 \\ x_1 + x_2 &\leqq 4. \end{aligned}$$

Solution. By means of slack variables we achieve the normal form of the constraints

$$\begin{aligned} z - 2x_1 - x_2 \qquad\qquad\qquad &= 0 \\ x_1 - \tfrac{1}{2}x_2 - x_3 \qquad\qquad &= 1 \\ x_1 - x_2 \qquad + x_4 \qquad &= 2 \\ x_1 + x_2 \qquad\qquad + x_5 &= 4 \\ x_i \geqq 0 \quad (i = 1, \cdots, 5). \end{aligned} \tag{8}$$

Note that the first slack variable is negative (or zero), which makes x_3 nonnegative within the feasibility region (and negative outside). From (7) and (8) we obtain the simplex table

$$
\begin{array}{c|cc|ccc|c}
z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\
\hline
1 & -2 & -1 & 0 & 0 & 0 & 0 \\
\hline
0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 \\
0 & 1 & -1 & 0 & 1 & 0 & 2 \\
0 & 1 & 1 & 0 & 0 & 1 & 4
\end{array}.
$$

x_1, x_2 are nonbasic, and we would like to take x_3, x_4, x_5 as basic variables. By our usual process of equating the nonbasic variables to zero we obtain from this table

$$x_1 = 0, \qquad x_2 = 0, \qquad x_3 = 1/(-1) = -1, \qquad x_4 = \tfrac{2}{1} = 2, \qquad x_5 = \tfrac{4}{1} = 4, \qquad z = 0.$$

$x_3 < 0$ indicates that (0, 0) lies outside the feasibility region. Since $x_3 < 0$, we cannot proceed immediately. Now, instead of searching for other basic variables, we use the following idea. Solving the second equation in (8) for x_3, we have

$$x_3 = -1 + x_1 - \tfrac{1}{2}x_2.$$

To this we now add a variable x_6 on the right,

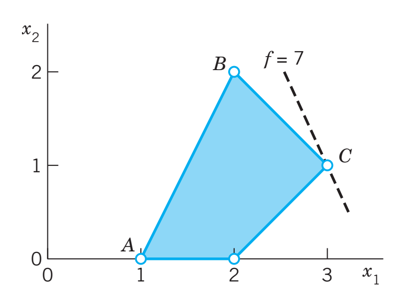


Fig. 476. Feasibility region in Example 2

$$x_3 = -1 + x_1 - \tfrac{1}{2}x_2 + x_6. \tag{9}$$

x_6 is called an **artificial variable** and is subject to the constraint $x_6 \geqq 0$.

We must take care that x_6 (which is not part of the given problem!) will disappear eventually. We shall see that we can accomplish this by adding a term $-Mx_6$ with very large M to the objective function. Because of (7) and (9) (solved for x_6) this gives the modified objective function for this "**extended problem**"

$$\hat{z} = z - Mx_6 = 2x_1 + x_2 - Mx_6 = (2 + M)x_1 + (1 - \tfrac{1}{2}M)x_2 - Mx_3 - M. \tag{10}$$

We see that the simplex table corresponding to (10) and (8) is

$$
\mathbf{T}_0 = \begin{array}{c|cc|cccc|c}
\hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \\
\hline
1 & -2 - M & -1 + \frac{1}{2}M & M & 0 & 0 & 0 & -M \\
\hline
0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\
0 & 1 & -1 & 0 & 1 & 0 & 0 & 2 \\
0 & 1 & 1 & 0 & 0 & 1 & 0 & 4 \\
0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 & 1
\end{array}.
$$

The last row of this table results from (9) written as $x_1 - \frac{1}{2}x_2 - x_3 + x_6 = 1$. We see that we can now start, taking x_4, x_5, x_6 as the basic variables and x_1, x_2, x_3 as the nonbasic variables. Column 2 has a negative first entry. We can take the second entry (1 in Row 2) as the pivot. This gives

$$
\mathbf{T}_1 = \begin{array}{c|cc|cccc|c}
\hat{z} & x_1 & x_2 & x_3 & x_4 & x_5 & x_6 & b \\
\hline
1 & 0 & -2 & -2 & 0 & 0 & 0 & 2 \\
\hline
0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 0 & 1 \\
0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 0 & 1 \\
0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 0 & 3 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0
\end{array}.
$$

This corresponds to $x_1 = 1, x_2 = 0$ (point A in Fig. 476), $x_3 = 0, x_4 = 1, x_5 = 3, x_6 = 0$. We can now drop Row 5 and Column 7. In this way we get rid of x_6, as wanted, and obtain

$$
\mathbf{T}_2 = \begin{array}{c|cc|ccc|c}
z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\
\hline
1 & 0 & -2 & -2 & 0 & 0 & 2 \\
\hline
0 & 1 & -\frac{1}{2} & -1 & 0 & 0 & 1 \\
0 & 0 & -\frac{1}{2} & 1 & 1 & 0 & 1 \\
0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3
\end{array}.
$$

In Column 3 we choose $\frac{3}{2}$ as the next pivot. We obtain

$$
\mathbf{T}_3 = \begin{array}{c|cc|ccc|c}
z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\
\hline
1 & 0 & 0 & -\frac{2}{3} & 0 & \frac{4}{3} & 6 \\
\hline
0 & 1 & 0 & -\frac{2}{3} & 0 & \frac{1}{3} & 2 \\
0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\
0 & 0 & \frac{3}{2} & 1 & 0 & 1 & 3
\end{array}.
$$

This corresponds to $x_1 = 2, x_2 = 2$ (this is B in Fig. 476), $x_3 = 0, x_4 = 2, x_5 = 0$. In Column 4 we choose $\frac{4}{3}$ as the pivot, by the usual principle. This gives

$$
\mathbf{T}_4 = \begin{array}{c|cc|ccc|c}
z & x_1 & x_2 & x_3 & x_4 & x_5 & b \\
\hline
1 & 0 & 0 & 0 & \frac{1}{2} & \frac{3}{2} & 7 \\
\hline
0 & 1 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 3 \\
0 & 0 & 0 & \frac{4}{3} & 1 & \frac{1}{3} & 2 \\
0 & 0 & \frac{3}{2} & 0 & -\frac{3}{4} & \frac{3}{4} & \frac{3}{2}
\end{array}.
$$

This corresponds to $x_1 = 3, x_2 = 1$ (point C in Fig. 476), $x_3 = \frac{3}{2}, x_4 = 0, x_5 = 0$. This is the maximum $f_{\max} = f(3, 1) = 7$. ■

We have reached the end of our discussion on linear programming. We have presented the simplex method in great detail as this method has many beautiful applications and works well on most practical problems. Indeed, problems of optimization appear in civil engineering, chemical engineering, environmental engineering, management science, logistics, strategic planning, operations management, industrial engineering, finance, and other areas. Furthermore, the simplex method allows your problem to be *scaled up* from a small modeling attempt to a larger modeling attempt, by adding more constraints and

variables, thereby making your model more realistic. The area of optimization is an active field of development and research and optimization methods, besides the simplex method, are being explored and experimented with.

PROBLEM SET 22.4

1. Maximize $z = f_1(\mathbf{x}) = 7x_1 + 14x_2$ subject to $0 \leqq x_1 \leqq 6$, $0 \leqq x_2 \leqq 3$, $7x_1 + 14x_2 \leqq 84$.
2. Do Prob. 1 with the last two constraints interchanged.
3. Maximize the daily output in producing x_1 steel sheets by process P_A and x_2 steel sheets by process P_B subject to the constraints of labor hours, machine hours, and raw material supply:
$$3x_1 + 2x_2 \leqq 180, \qquad 4x_1 + 6x_2 \leqq 200,$$
$$5x_1 + 3x_2 \leqq 160.$$
4. Maximize $z = 300x_1 + 500x_2$ subject to $2x_1 + 8x_2 \leqq 60$, $2x_1 + x_2 \leqq 30$, $4x_1 + 4x_2 \leqq 60$.
5. Do Prob. 4 with the last two constraints interchanged. Comment on the resulting simplification.
6. Maximize the total output $f = x_1 + x_2 + x_3$ (production from three distinct processes) subject to input constraints (limitation of time available for production)
$$5x_1 + 6x_2 + 7x_3 \leqq 12,$$
$$7x_1 + 4x_2 + x_3 \leqq 12.$$
7. Maximize $f = 5x_1 + 8x_2 + 4x_3$ subject to $x_j \geqq 0$ $(j = 1, \cdots, 5)$ and $x_1 + x_3 + x_5 = 1$, $x_2 + x_3 + x_4 = 1$.
8. Using an artificial variable, minimize $f = 4x_1 - x_2$ subject to $x_1 + x_2 \geqq 2$, $-2x_1 + 3x_2 \leqq 1$, $5x_1 + 4x_2 \leqq 50$.
9. Maximize $f = 2x_1 + 3x_2 + 2x_3$, $x_1 \geqq 0$, $x_2 \geqq 0$, $x_3 \geqq 0$, $x_1 + 2x_2 - 4x_3 \leqq 2$, $x_1 + 2x_2 + 2x_3 \leqq 5$.

CHAPTER 22 REVIEW QUESTIONS AND PROBLEMS

1. What is unconstrained optimization? Constraint optimization? To which one do methods of calculus apply?
2. State the idea and the formulas of the method of steepest descent.
3. Write down an algorithm for the method of steepest descent.
4. Design a "method of steepest ascent" for determining maxima.
5. What is the method of steepest descent for a function of a single variable?
6. What is the basic idea of linear programming?
7. What is an objective function? A feasible solution?
8. What are slack variables? Why did we introduce them?
9. What happens in Example 1 of Sec. 22.1 if you replace $f(\mathbf{x}) = x_1^2 + 3x_2^2$ with $f(\mathbf{x}) = x_1^2 + 5x_2^2$? Start from $\mathbf{x}_0 = [6 \quad 3]^\mathsf{T}$. Do 5 steps. Is the convergence faster or slower?
10. Apply the method of steepest descent to $f(\mathbf{x}) = 9x_1^2 + x_2^2 + 18x_1 - 4x_2$, 5 steps. Start from $\mathbf{x}_0 = [2 \quad 4]^\mathsf{T}$.
11. In Prob. 10, could you start from $[0 \quad 0]^\mathsf{T}$ and do 5 steps?
12. Show that the gradients in Prob. 11 are orthogonal. Give a reason.

13–16 Graph or sketch the region in the first quadrant of the x_1x_2-plane determined by the following inequalities.

13. $x_1 - 2x_2 \leqq -2$, $\quad 0.8x_1 + x_2 \leqq 6$
14. $x_1 - 2x_2 \geqq -4$, $\quad 2x_1 + x_2 \leqq 12$, $\quad x_1 + x_2 \leqq 8$
15. $x_1 + x_2 \leqq 5$, $\quad x_2 \leqq 3$, $\quad -x_1 + x_2 \leqq 2$
16. $x_1 + x_2 \geqq 2$, $\quad 2x_1 - 3x_2 \geqq -12$, $\quad x_1 \leqq 15$

17–20 Maximize or minimize as indicated.

17. Maximize $f = 10x_1 + 20x_2$ subject to $x_1 \leqq 5$, $x_1 + x_2 \leqq 6$, $x_2 \leqq 4$.
18. Maximize $f = x_1 + x_2$ subject to $x_1 + 2x_2 \leqq 10$, $2x_2 + x_2 \leqq 10$, $x_2 \leqq 4$.
19. Minimize $f = 2x_1 - 10x_2$ subject to $x_1 - x_2 \leqq 4$, $2x_1 + x_2 \leqq 14$, $x_1 + x_2 \leqq 9$, $-x_1 + 3x_2 \leqq 15$.
20. A factory produces two kinds of gaskets, G_1, G_2, with net profit of \$60 and \$30, respectively, Maximize the total daily profit subject to the constraints (x_j = number of gaskets G_j produced per day):
$$40x_1 + 40x_2 \leqq 1800 \quad \text{(Machine hours)},$$
$$200x_1 + 20x_2 \leqq 6300 \quad \text{(Labor)}.$$

SUMMARY OF CHAPTER 22

Unconstrained Optimization. Linear Programming

In optimization problems we maximize or minimize an ***objective function*** $z = f(\mathbf{x})$ depending on control variables $x_1, \cdots, x_m$ whose domain is either unrestricted ("***unconstrained optimization***," Sec. 22.1) or restricted by constraints in the form of inequalities or equations or both ("***constrained optimization***," Sec. 22.2).

If the objective function is *linear* and the constraints are *linear inequalities* in $x_1, \cdots, x_m$, then by introducing **slack variables** $x_{m+1}, \cdots, x_n$ we can write the optimization problem in **normal form** with the objective function given by

$$f_1 = c_1x_1 + \cdots + c_nx_n \tag{1}$$

(where $c_{m+1} = \cdots = c_n = 0$) and the constraints given by

$$\begin{gathered} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n = b_1 \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n = b_m \\ x_1 \geqq 0, \cdots, x_n \geqq 0. \end{gathered} \tag{2}$$

In this case we can then apply the widely used ***simplex method*** (Sec. 22.3), a systematic stepwise search through a very much reduced subset of all feasible solutions. Section 22.4 shows how to overcome difficulties with this method.

Graphs. Combinatorial Optimization

Many problems in electrical engineering, civil engineering, operations research, industrial engineering, management, logistics, marketing, and economics can be modeled by ***graphs*** and directed ***graphs***, called digraphs. This is not surprising as they allow us to model networks, such as roads and cables, where the nodes may be cities or computers. The task then is to find the shortest path through the network or the best way to connect computers. Indeed, many researchers who made contributions to combinatorial optimization and graphs, and whose names lend themselves to fundamental algorithms in this chapter, such as Fulkerson, Kruskal, Moore, and Prim, all worked at Bell Laboratories in New Jersey, the major R&D facilities of the huge telephone and telecommunication company AT&T. As such, they were interested in methods of optimally building computer networks and telephone networks. The field has progressed into looking for more and more efficient algorithms for very large problems.

Combinatorial optimization deals with optimization problems that are of a pronounced discrete or combinatorial nature. Often the problems are very large and so a direct search may not be possible. Just like in linear programming (Chap. 22), the computer is an indispensible tool and makes solving large-scale modeling problems possible. Because the area has a distinct flavor, different from ODEs, linear algebra, and other areas, we start with the basics and gradually introduce algorithms for shortest path problems (Secs. 22.2, 22.3), shortest spanning trees (Secs. 23.4, 23.5), flow problems in networks (Secs. 23.6, 23.7), and assignment problems (Sec. 23.8).

Prerequisite: none.
References and Answers to Problems: App. 1 Part F, App. 2.

23.1 Graphs and Digraphs

Roughly, a *graph* consists of points, called *vertices*, and lines connecting them, called *edges*. For example, these may be four cities and five highways connecting them, as in Fig. 477. Or the points may represent some people, and we connect by an edge those who do business with each other. Or the vertices may represent computers in a network and the edge connections between them. Let us now give a formal definition.

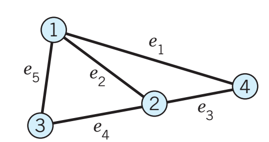


Fig. 477. Graph consisting of 4 vertices and 5 edges

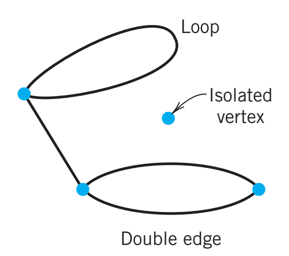


Fig. 478. Isolated vertex, loop, double edge. (Excluded by definition.)

DEFINITION

Graph

A **graph** G consists of two finite sets (sets having finitely many elements), a set V of points, called **vertices**, and a set E of connecting lines, called **edges**, such that each edge connects two vertices, called the *endpoints* of the edge. We write

$$G = (V, E).$$

Excluded are *isolated vertices* (vertices that are not endpoints of any edge), *loops* (edges whose endpoints coincide), and *multiple edges* (edges that have both endpoints in common). See Fig. 478.

CAUTION! Our three exclusions are practical and widely accepted, but not uniformly. For instance, some authors permit multiple edges and call graphs without them *simple graphs.* ■

We denote vertices by letters, $u, v, \cdots$ or $v_1, v_2, \cdots$ or simply by numbers $1, 2, \cdots$ (as in Fig. 477). We denote edges by $e_1, e_2, \cdots$ or by their two endpoints; for instance, $e_1 = (1, 4)$, $e_2 = (1, 2)$ in Fig. 477.

An edge (v_i, v_j) is called **incident** with the vertex v_i (and conversely); similarly, (v_i, v_j) is *incident* with v_j. The number of edges incident with a vertex v is called the **degree** of v. Two vertices are called **adjacent** in G if they are connected by an edge in G (that is, if they are the two endpoints of some edge in G).

We meet graphs in different fields under different names: as "networks" in electrical engineering, "structures" in civil engineering, "molecular structures" in chemistry, "organizational structures" in economics, "sociograms," "road maps," "telecommunication networks," and so on.

Digraphs (Directed Graphs)

Nets of one-way streets, pipeline networks, sequences of jobs in construction work, flows of computation in a computer, producer–consumer relations, and many other applications suggest the idea of a "digraph" (= directed graph), in which each edge has a direction (indicated by an arrow, as in Fig. 479).

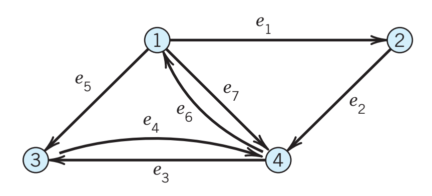


Fig. 479. Digraph

DEFINITION

Digraph (Directed Graph)

A **digraph** $G = (V, E)$ is a graph in which each edge $e = (i, j)$ has a direction from its "*initial point*" i to its "*terminal point*" j.

Two edges connecting the same two points i, j are now permitted, provided they have opposite directions, that is, they are (i, j) and (j, i). *Example.* (1, 4) and (4, 1) in Fig. 479.

A **subgraph** or subdigraph of a given graph or digraph $G = (V, E)$, respectively, is a graph or digraph obtained by deleting some of the edges and vertices of G, retaining the other edges of G (together with their pairs of endpoints). For instance, e_1, e_3 (together with the vertices 1, 2, 4) form a subgraph in Fig. 477, and e_3, e_4, e_5 (together with the vertices 1, 3, 4) form a subdigraph in Fig. 479.

Computer Representation of Graphs and Digraphs

Drawings of graphs are useful to people in explaining or illustrating specific situations. Here one should be aware that a graph may be sketched in various ways; see Fig. 480. For handling graphs and digraphs in computers, one uses matrices or lists as appropriate data structures, as follows.

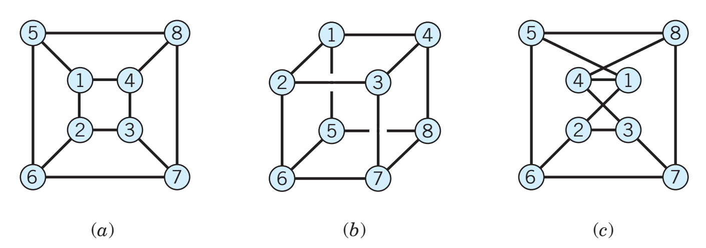


Fig. 480. Different sketches of the same graph

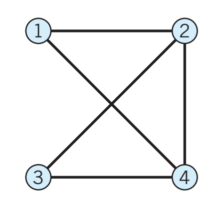
Adjacency Matrix of a Graph G: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has an edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

Thus $a_{ij} = 1$ if and only if two vertices i and j are adjacent in G. Here, by definition, no vertex is considered to be adjacent to itself; thus, $a_{ii} = 0$. $\mathbf{A}$ is symmetric, $a_{ij} = a_{ji}$. (Why?)

The adjacency matrix of a graph is generally much smaller than the so-called *incidence matrix* (see Prob. 18) and is preferred over the latter if one decides to store a graph in a computer in matrix form.

EXAMPLE 1 **Adjacency Matrix of a Graph**



$$\begin{array}{r} \text{Vertex} \\ \text{Vertex } 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix} \end{array}$$

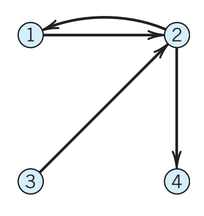
■

Adjacency Matrix of a Digraph *G*: Matrix $\mathbf{A} = [a_{ij}]$ with entries

$$a_{ij} = \begin{cases} 1 & \text{if } G \text{ has a directed edge } (i, j), \\ 0 & \text{else.} \end{cases}$$

This matrix **A** need not be symmetric. (Why?)

EXAMPLE 2 **Adjacency Matrix of a Digraph**



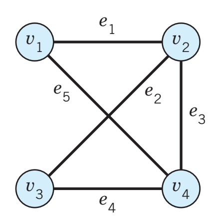
$$\begin{array}{r} \text{To vertex} \\ \text{From vertex } 1 \\ 2 \\ 3 \\ 4 \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & 3 & 4 \end{array} \\ \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

■

Lists. The **vertex incidence list** of a graph shows, for each vertex, the incident edges. The **edge incidence list** shows for each edge its two endpoints. Similarly for a *digraph*; in the vertex list, outgoing edges then get a minus sign, and in the edge list we now have *ordered* pairs of vertices.

EXAMPLE 3 **Vertex Incidence List and Edge Incidence List of a Graph**

This graph is the same as in Example 1, except for notation.



Vertex	Incident Edges
v_1	e_1, e_5
v_2	e_1, e_2, e_3
v_3	e_2, e_4
v_4	e_3, e_4, e_5

Edge	Endpoints
e_1	v_1, v_2
e_2	v_2, v_3
e_3	v_2, v_4
e_4	v_3, v_4
e_5	v_1, v_4

■

Sparse graphs are graphs with few edges (far fewer than the maximum possible number $n(n-1)/2$, where n is the number of vertices). For these graphs, matrices are not efficient. *Lists* then have the advantage of requiring much less storage and being easier to handle; they can be ordered, sorted, or manipulated in various other ways directly within the computer. For instance, in tracing a "walk" (a connected sequence of edges with pairwise common endpoints), one can easily go back and forth between the two lists just discussed, instead of scanning a large column of a matrix for a single 1.

Computer science has developed more refined lists, which, in addition to the actual content, contain "pointers" indicating the preceding item or the next item to be scanned or both items (in the case of a "walk": the preceding edge or the subsequent one). For details, see Refs. [E16] and [F7].

This section was devoted to basic concepts and notations needed throughout this chapter, in which we shall discuss some of the most important classes of combinatorial optimization problems. This will at the same time help us to become more and more familiar with graphs and digraphs.

PROBLEM SET 23.1

1. Explain how the following can be regarded as a graph or a digraph: a family tree, air connections between given cities, trade relations between countries, a tennis tournament, and memberships of some persons in some committees.

2. Sketch the graph consisting of the vertices and edges of a triangle. Of a pentagon. Of a tetrahedron.

3. How would you represent a net of two-way and one-way streets by a digraph?

4. Worker W_1 can do jobs J_1, J_3, J_4, worker W_2 job J_3, and worker W_3 jobs J_2, J_3, J_4. Represent this by a graph.

5. Find further situations that can be modeled by a graph or diagraph.

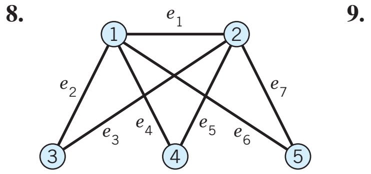
ADJACENCY MATRIX

6. Show that the adjacency matrix of a graph is symmetric.

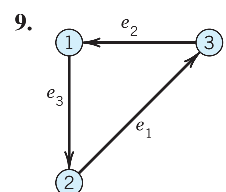
7. When will the adjacency matrix of a digraph be symmetric?

8–13 Find the adjacency matrix of the given graph or digraph.

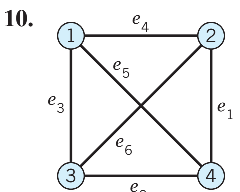
8.



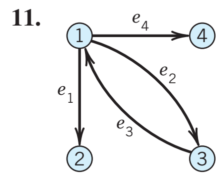
9.



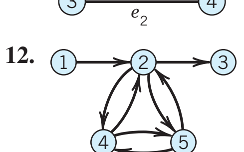
10.



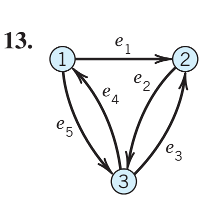
11.



12.



13.



14–15 Sketch the graph for the given adjacency matrix.

14. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$ 15. $\begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{bmatrix}$

16. **Complete graph.** Show that a graph G with n vertices can have at most $n(n-1)/2$ edges, and G has exactly $n(n-1)/2$ edges if G is *complete*, that is, if every pair of vertices of G is joined by an edge. (Recall that loops and multiple edges are excluded.)

17. In what case are all the off-diagonal entries of the adjacency matrix of a graph G equal to one?

18. Incidence matrix B of a graph. The definition is $\mathbf{B} = [b_{jk}]$, where

$$b_{jk} = \begin{cases} 1 & \text{if vertex } j \text{ is an endpoint of edge } e_k, \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of the graph in Prob. **8**.

19. Incidence matrix $\widetilde{\mathbf{B}}$ of a digraph. The definition is $\widetilde{\mathbf{B}} = [\widetilde{b}_{jk}]$, where

$$\widetilde{b}_{jk} = \begin{cases} -1 & \text{if edge } e_k \text{ leaves vertex } j, \\ 1 & \text{if edge } e_k \text{ enters vertex } j, \\ 0 & \text{otherwise.} \end{cases}$$

Find the incidence matrix of the digraph in Prob. 11.

20. Make the vertex incidence list of the digraph in Prob. 11.

23.2 Shortest Path Problems. Complexity

The rest of this chapter is devoted to the most important classes of problems of combinatorial optimization that can be represented by graphs and digraphs. We selected these problems because of their importance in applications, and present their solutions in algorithmic form. Although basic ideas and algorithms will be explained and illustrated by small graphs, you should keep in mind that real-life problems may often involve many thousands or even millions of vertices and edges. Think of computer networks, telephone networks, electric power grids, worldwide air travel, and companies that have offices and stores in all larger cities. You can also think of other ideas for networks related to the Internet, such as electronic commerce (networks of buyers and sellers of goods over the Internet) and social networks and related websites, such as Facebook. Hence reliable and efficient systematic methods are an absolute necessity—solutions by trial and error would no longer work, even if "nearly optimal" solutions were acceptable.

We begin with **shortest path problems**, as they arise, for instance, in designing shortest (or least expensive, or fastest) routes for a traveling salesman, for a cargo ship, etc. Let us first explain what we mean by a path.

In a graph $G = (V, E)$ we can walk from a vertex v_1 along some edges to some other vertex v_k. Here we can

(A) make no restrictions, or

(B) require that each *edge* of G be traversed at most once, or

(C) require that each *vertex* be visited at most once.

In case (A) we call this a **walk**. Thus a walk from v_1 to v_k is of the form

$$(1) \qquad (v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k),$$

where some of these edges or vertices may be the same. In case (B), where each *edge* may occur at most once, we call the walk a **trail**. Finally, in case (C), where each *vertex* may occur at most once (and thus each edge automatically occurs at most once), we call the trail a **path**.

We admit that a walk, trail, or path may end at the vertex it started from, in which case we call it **closed**; then $v_k = v_1$ in (1).

A closed path is called a **cycle**. *A cycle has at least three edges* (because we do not have double edges; see Sec. 23.1). Figure 481 illustrates all these concepts.

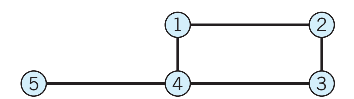


Fig. 481. Walk, trail, path, cycle

1 − 2 − 3 − 2 is a walk (not a trail).
4 − 1 − 2 − 3 − 4 − 5 is a trail (not a path).
1 − 2 − 3 − 4 − 5 is a path (not a cycle).
1 − 2 − 3 − 4 − 1 is a cycle.

Shortest Path

To define the concept of a shortest path, we assume that $G = (V, E)$ is a **weighted graph**, that is, each edge (v_i, v_j) in G has a given *weight* or *length* $l_{ij} > 0$. Then a **shortest path** $v_1 \rightarrow v_k$ (with fixed v_1 and v_k) is a path (1) such that the sum of the lengths of its edges

$$l_{12} + l_{23} + l_{34} + \cdots + l_{k-1,k}$$

(l_{12} = length of (v_1, v_2), etc.) is minimum (as small as possible among all paths from v_1 to v_k). Similarly, a **longest path** $v_1 \rightarrow v_k$ is one for which that sum is maximum.

Shortest (and longest) path problems are among the most important optimization problems. Here, "length" l_{ij} (often also called "cost" or "weight") can be an actual length measured in miles or travel time or fuel expenses, but it may also be something entirely different.

For instance, the *traveling salesman problem* requires the determination of a shortest **Hamiltonian**[1] **cycle** in a graph, that is, a cycle that contains all the vertices of the graph.

In more detail, the traveling salesman problem in its most basic and intuitive form can be stated as follows. You have a salesman who has to drive by car to his customers. He has to drive to n cities. He can start at any city and after completion of the trip he has to return to that city. Furthermore, he can only visit each city once. All the cities are linked by roads to each other, so any city can be visited from any other city directly, that is, if he wants to go from one city to another city, there is only one direct road connecting those two cities. He has to find the optimal route, that is, the route with the shortest total mileage for the overall trip. This is a classic problem in combinatorial optimization and comes up in many different versions and applications. The maximum number of possible paths to be examined in the process of selecting the optimal path for n cities is $(n - 1)!/2$, because, after you pick the first city, you have $n - 1$ choices for the second city, $n - 2$ choices for the third city, etc. You get a total of $(n - 1)!$ (see Sec. 24.4). However, since the mileage does not depend on the direction of the tour (e.g., for $n = 4$ (four cities 1, 2, 3, 4), the tour 1–2–3–4–1 has the same mileage as 1–4–3–2–1, etc., so that we counted all the tours twice!), the final answer is $(n - 1)!/2$. Even for a small number of cities, say $n = 15$, the maximum number of possible paths is very large. Use your calculator or CAS to see for yourself! This means that this is a very difficult problem for larger n and typical of problems in combinatorial optimization, in that you want a discrete solution but where it might become nearly impossible to explicitly search through all the possibilities and therefore some heuristics (rules of thumbs, shortcuts) might be used, and a less than optimal answer suffices.

[1]WILLIAM ROWAN HAMILTON (1805–1865), Irish mathematician, known for his work in dynamics.

A variation of the traveling salesman problem is the following. By choosing the "most profitable" route $v_1 \to v_k$, a salesman may want to maximize Σl_{ij}, where l_{ij} is his expected commission minus his travel expenses for going from town i to town j.

In an investment problem, i may be the day an investment is made, j the day it matures, and l_{ij} the resulting profit, and one gets a graph by considering the various possibilities of investing and reinvesting over a given period of time.

Shortest Path If All Edges Have Length $l = 1$

Obviously, if all edges have length l, then a shortest path $v_1 \to v_k$ is one that has the smallest number of edges among all paths $v_1 \to v_k$ in a given graph G. For this problem we discuss a BFS algorithm. BFS stands for **Breadth First Search**. This means that in each step the algorithm visits *all neighboring* (all adjacent) vertices of a vertex reached, as opposed to a DFS algorithm (**Depth First Search** algorithm), which makes a long trail (as in a maze). This widely used BFS algorithm is shown in Table 23.1.

We want to find a shortest path in G from a vertex s (***start***) to a vertex t (***terminal***). To guarantee that there is a path from s to t, we make sure that G does not consist of separate portions. Thus we assume that G is **connected**, that is, for any two vertices v and w there is a path $v \to w$ in G. (Recall that a vertex v is called **adjacent** to a vertex u if there is an edge (u, v) in G.)

Table 23.1 Moore's[2] BFS for Shortest Path (All Lengths One)

Proceedings of the International Symposium for Switching Theory, Part II. pp. 285–292. Cambridge: Harvard University Press, 1959.

ALGORITHM MOORE $[G = (V, E), s, t]$

This algorithm determines a shortest path in a connected graph $G = (V, E)$ from a vertex s to a vertex t.

INPUT: Connected graph $G = (V, E)$, in which one vertex is denoted by s and one by t, and each edge (i, j) has length $l_{ij} = 1$. Initially all vertices are unlabeled.

OUTPUT: A shortest path $s \to t$ in $G = (V, E)$

1. Label s with 0.
2. Set $i = 0$.
3. Find all *unlabeled* vertices adjacent to a vertex labeled i.
4. Label the vertices just found with $i + 1$.
5. If vertex t is labeled, then "backtracking" gives the shortest path

$$k\ (= \text{label of } t),\ k - 1,\ k - 2, \cdots, 0$$

OUTPUT $k, k - 1, k - 2, \cdots, 0$. Stop

Else increase i by 1. Go to Step 3.

End MOORE

[2]EDWARD FORREST MOORE (1925–2003), American mathematician and computer scientist, who did pioneering work in theoretical computer science (automata theory, Turing machines).

EXAMPLE 1 Application of Moore's BFS Algorithm

Find a shortest path $s \rightarrow t$ in the graph G shown in Fig. 482.

Solution. Figure 482 shows the labels. The blue edges form a shortest path (length 4). There is another shortest path $s \rightarrow t$. (Can you find it?) Hence in the program we must introduce a rule that makes backtracking unique because otherwise the computer would not know what to do next if at some step there is a choice (for instance, in Fig. 482 when it got back to the vertex labeled 2). The following rule seems to be natural.

Backtracking rule. Using the numbering of the vertices from 1 to n (not the labeling!), at each step, if a vertex labeled i is reached, take as the next vertex that with the smallest number (not label!) among all the vertices labeled $i - 1$. ■

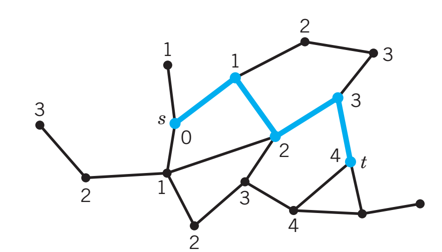


Fig. 482. Example 1, given graph and result of labeling

Complexity of an Algorithm

Complexity *of Moore's algorithm.* To find the vertices to be labeled 1, we have to scan all edges incident with s. Next, when $i = 1$, we have to scan all edges incident with vertices labeled 1, etc. Hence each edge is scanned twice. These are $2m$ operations (m = number of edges of G). This is a function $c(m)$. Whether it is $2m$ or $5m + 3$ or $12m$ is not so essential; it *is* essential that $c(m)$ is proportional to m (not m^2, for example); it is of the "order" m. We write for any function $am + b$ simply $O(m)$, for any function $am^2 + bm + d$ simply $O(m^2)$, and so on; here, O suggests **order**. The underlying idea and practical aspect are as follows.

In judging an algorithm, we are mostly interested in its behavior for very large problems (large m in the present case), since these are going to determine the limits of the applicability of the algorithm. Thus, the essential item is the fastest growing term (am^2 in $am^2 + bm + d$, etc.) since it will overwhelm the others when m is large enough. Also, a constant factor in this term is not very essential; for instance, the difference between two algorithms of orders, say, $5m^2$ and $8m^2$ is generally not very essential and can be made irrelevant by a modest increase in the speed of computers. However, it does make a great practical difference whether an algorithm is of order m or m^2 or of a still higher power m^p. And the biggest difference occurs between these "polynomial orders" and "exponential orders," such as 2^m.

For instance, on a computer that does 10^9 operations per second, a problem of size $m = 50$ will take 0.3 sec with an algorithm that requires m^5 operations, but 13 days with an algorithm that requires 2^m operations. But this is not our only reason for regarding polynomial orders as good and exponential orders as bad. Another reason is the ***gain in using a faster computer***. For example, let two algorithms be $O(m)$ and $O(m^2)$. Then, since $1000 = 31.6^2$, an increase in speed by a factor 1000 has the effect that per hour we can do problems 1000 and 31.6 times as big, respectively. But since $1000 = 2^{9.97}$, with an algorithm that is $O(2^m)$, all we gain is a relatively modest increase of 10 in problem size because $2^{9.97} \cdot 2^m = 2^{m+9.97}$.

The **symbol *O*** is quite practical and commonly used whenever the order of growth is essential, but not the specific form of a function. Thus if a function $g(m)$ is of the form

$$g(m) = kh(m) + \text{more slowly growing terms} \qquad (k \neq 0, \text{ constant}),$$

we say that $g(m)$ is of the *order* $h(m)$ and write

$$g(m) = O(h(m)).$$

For instance,

$$am + b = O(m), \qquad am^2 + bm + d = O(m^2), \qquad 5 \cdot 2^m + 3m^2 = O(2^m).$$

We want an algorithm $\mathcal{A}$ to be "efficient," that is, "good" with respect to

(i) *Time* (number $c_{\mathcal{A}}(m)$ of computer operations), or

(ii) *Space* (storage needed in the internal memory)

or both. Here $c_{\mathcal{A}}$ suggests "**complexity**" of $\mathcal{A}$. Two popular choices for $c_{\mathcal{A}}$ are

(*Worst case*) $c_{\mathcal{A}}(m)$ = longest time $\mathcal{A}$ takes for a problem of size m,

(*Average case*) $c_{\mathcal{A}}(m)$ = average time $\mathcal{A}$ takes for a problem of size m.

In problems on graphs, the "size" will often be m (number of edges) or n (number of vertices). For Moore's algorithm, $c_{\mathcal{A}}(m) = 2m$ in both cases. Hence the complexity of Moore's algorithm is of order $O(m)$.

For a "good" algorithm $\mathcal{A}$, we want that $c_{\mathcal{A}}(m)$ does not grow too fast. Accordingly, we call $\mathcal{A}$ **efficient** if $c_{\mathcal{A}}(m) = O(m^k)$ for some integer $k \geqq 0$; that is, $c_{\mathcal{A}}$ may contain only powers of m (or functions that grow even more slowly, such as ln m), but no exponential functions. Furthermore, we call $\mathcal{A}$ **polynomially bounded** if $\mathcal{A}$ is efficient when we choose the "worst case" $c_{\mathcal{A}}(m)$. These conventional concepts have intuitive appeal, as our discussion shows.

Complexity should be investigated for every algorithm, so that one can also compare different algorithms for the same task. This may often exceed the level in this chapter; accordingly, we shall confine ourselves to a few occasional comments in this direction.

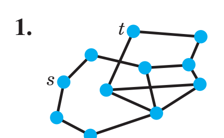
PROBLEM SET 23.2

SHORTEST PATHS, MOORE'S BFS

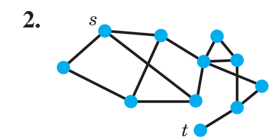
(All edges length one)

1–4 Find a shortest path $P: s \to t$ and its length by Moore's algorithm. Sketch the graph with the labels and indicate P by heavier lines as in Fig. 482.

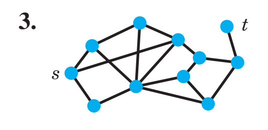
1.



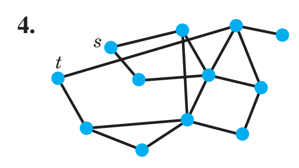
2.



3.



4.



5. Moore's algorithm. Show that if vertex v has label $\lambda(v) = k$, then there is a path $s \to v$ of length k.

6. Maximum length. What is the maximum number of edges that a shortest path between any two vertices in a graph with n vertices can have? Give a reason. In a complete graph with all edges of length 1?

7. Nonuniqueness. Find another shortest path from s to t in Example 1 of the text.

8. Moore's algorithm. Call the length of a shortest path $s \rightarrow v$ the *distance* of v from s. Show that if v has distance l, it has label $\lambda(v) = l$.

9. CAS PROBLEM. Moore's Algorithm. Write a computer program for the algorithm in Table 23.1. Test the program with the graph in Example 1. Apply it to Probs. 1–3 and to some graphs of your own choice.

10–12 HAMILTONIAN CYCLE

10. Find and sketch a Hamiltonian cycle in the graph of a dodecahedron, which has 12 pentagonal faces and 20 vertices (Fig. 483). This is a problem Hamilton himself considered.

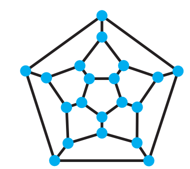


Fig. 483. Problem 10

11. Find and sketch a Hamiltonian cycle in Prob. 1.

12. Does the graph in Prob. 4 have a Hamiltonian cycle?

13–14 POSTMAN PROBLEM

13. The **postman problem** is the problem of finding a closed walk $W: s \rightarrow s$ (s the post office) in a graph G with edges (i, j) of length $l_{ij} > 0$ such that every edge of G is traversed at least once and the length of W is minimum. Find a solution for the graph in Fig. 484 by inspection. (The problem is also called the *Chinese postman problem* since it was published in the journal *Chinese Mathematics* 1 (1962), 273–277.)

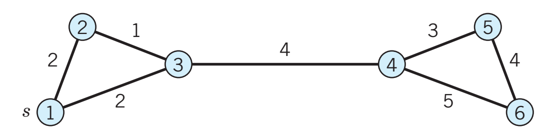


Fig. 484. Problem 13

14. Show that the length of a shortest postman trail is the same for every starting vertex.

15–17 EULER GRAPHS

15. An **Euler graph** G is a graph that has a closed Euler trail. An **Euler trail** is a trail that contains every edge of G exactly once. Which subgraph with four edges of the graph in Example 1, Sec. 23.1, is an Euler graph?

16. Find four different closed Euler trails in Fig. 485.

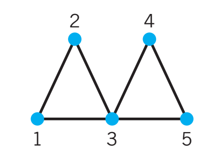


Fig. 485. Problem 16

17. Is the graph in Fig. 484 an Euler graph. Give reason.

18–20 ORDER

18. Show that $O(m^3) + O(m^3) = O(m^3)$ and $kO(m^p) = O(m^p)$.

19. Show that $\sqrt{1 + m^2} = O(m)$, $0.02e^m + 100m^2 = O(e^m)$.

20. If we switch from one computer to another that is 100 times as fast, what is our gain in problem size per hour in the use of an algorithm that is $O(m)$, $O(m^2)$, $O(m^5)$, $O(e^m)$?

23.3 Bellman's Principle. Dijkstra's Algorithm

We continue our discussion of the shortest path problem in a graph G. The last section concerned the special case that all edges had length 1. But in most applications the edges (i, j) will have any lengths $l_{ij} > 0$, and we now turn to this general case, which is of greater practical importance. We write $l_{ij} = \infty$ for any edge (i, j) that does not exist in G (setting $\infty + a = \infty$ for any number a, as usual).

We consider the problem of finding shortest paths from a given vertex, denoted by 1 and called the **origin**, to ***all*** other vertices $2, 3, \cdots, n$ of G. We let L_j denote the length of a shortest path P_j: $1 \rightarrow j$ in G.

THEOREM 1

Bellman's Minimality Principle or Optimality Principle[3]

If P_j: $1 \rightarrow j$ is a shortest path from 1 to j in G and (i, j) is the last edge of P_j (Fig. 486), *then P_i: $1 \rightarrow i$ [obtained by dropping (i, j) from P_j] is a shortest path $1 \rightarrow i$.*

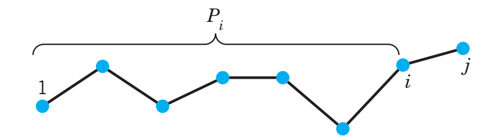


Fig. 486. Paths P and P_i in Bellman's minimality principle

PROOF Suppose that the conclusion is false. Then there is a path P_i^*: $1 \rightarrow i$ that is shorter than P_i. Hence, if we now add (i, j) to P_i^*, we get a path $1 \rightarrow j$ that is shorter than P_j. This contradicts our assumption that P_j is shortest. ∎

From Bellman's principle we can derive basic equations as follows. For fixed j we may obtain various paths $1 \rightarrow j$ by taking shortest paths P_i for various i for which there is in G an edge (i, j), and add (i, j) to the corresponding P_i. These paths obviously have lengths $L_i + l_{ij}$ (L_i = length of P_i). We can now take the minimum over i, that is, pick an i for which $L_i + l_{ij}$ is smallest. By the Bellman principle, this gives a shortest path $1 \rightarrow j$. It has the length

$$\begin{aligned} L_1 &= 0 \\ L_j &= \min_{i \neq j} (L_i + l_{ij}), \end{aligned} \qquad j = 2, \cdots, n. \tag{1}$$

These are the **Bellman equations**. Since $l_{ii} = 0$ by definition, instead of $\min_{i \neq j}$ we can simply write $\min_i$. These equations suggest the idea of one of the best-known algorithms for the shortest path problem, as follows.

Dijkstra's Algorithm for Shortest Paths

Dijkstra's[4] **algorithm** is shown in Table 23.2, where a **connected graph** G is a graph in which, for any two vertices v and w in G, there is a path $v \rightarrow w$. The algorithm is a labeling procedure. At each stage of the computation, each vertex v gets a label, either

(PL) a *permanent label* = length L_v of a shortest path $1 \rightarrow v$

or

(TL) a *temporary label* = upper bound $\tilde{L}_v$ for the length of a shortest path $1 \rightarrow v$.

[3]RICHARD BELLMAN (1920–1984), American mathematician, known for his work in dynamic programming.

[4]EDSGER WYBE DIJKSTRA (1930–2002), Dutch computer scientist, 1972 recipient of the ACM Turing Award. His algorithm appeared in *Numerische Mathematik* **1** (1959), 269–271.

We denote by $\mathcal{PL}$ and $\mathcal{TL}$ the sets of vertices with a permanent label and with a temporary label, respectively. The algorithm has an initial step in which vertex 1 gets the permanent label $L_1 = 0$ and the other vertices get temporary labels, and then the algorithm alternates between Steps 2 and 3. In Step 2 the idea is to pick k "minimally." In Step 3 the idea is that the upper bounds will in general improve (decrease) and must be updated accordingly. Namely, the new temporary label $\tilde{L}_j$ of vertex j will be the old one if there is no improvement or it will be $L_k + l_{kj}$ if there is.

Table 23.2 Dijkstra's Algorithm for Shortest Paths

ALGORITHM DIJKSTRA $[G = (V, E), V = \{1, \cdots, n\}, l_{ij}$ for all (i, j) in $E]$

Given a connected graph $G = (V, E)$ with vertices $1, \cdots, n$ and edges (i, j) having lengths $l_{ij} > 0$, this algorithm determines the lengths of shortest paths from vertex 1 to the vertices $2, \cdots, n$.

INPUT: Number of vertices n, edges (i, j), and lengths l_{ij}

OUTPUT: Lengths L_j of shortest paths $1 \to j$, $j = 2, \cdots, n$

1. *Initial step*

Vertex 1 gets PL: $L_1 = 0$.
Vertex j $(= 2, \cdots, n)$ gets TL: $\tilde{L}_j = l_{1j}$ $(= \infty$ if there is no edge $(1, j)$ in $G)$.
Set $\mathcal{PL} = \{1\}$, $\mathcal{TL} = \{2, 3, \cdots, n\}$.

2. *Fixing a permanent label*

Find a k in $\mathcal{TL}$ for which $\tilde{L}_k$ is miminum, set $L_k = \tilde{L}_k$. Take the smallest k if there are several. Delete k from $\mathcal{TL}$ and include it in $\mathcal{PL}$.
If $\mathcal{TL} = \varnothing$ (that is, $\mathcal{TL}$ is empty) then

OUTPUT $L_2, \cdots, L_n$. Stop

Else continue (that is, go to Step 3).

3. *Updating temporary labels*

For all j in $\mathcal{TL}$, set $\tilde{L}_j = \min_k \{\tilde{L}_j, L_k + l_{kj}\}$ (that is, take the smaller of $\tilde{L}_j$ and $L_k + l_{kj}$ as your new $\tilde{L}_j$).

Go to Step 2.

End DIJKSTRA

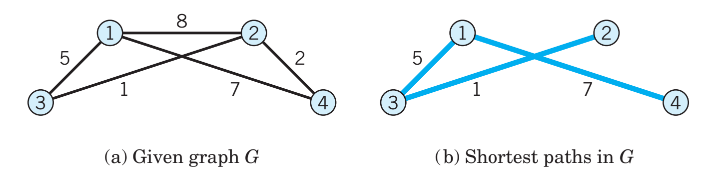
EXAMPLE 1 Application of Dijkstra's Algorithm

Applying Dijkstra's algorithm to the graph in Fig. 487a, find shortest paths from vertex 1 to vertices 2, 3, 4.

Solution. We list the steps and computations.

1.	$L_1 = 0, \tilde{L}_2 = 8, \tilde{L}_3 = 5, \tilde{L}_4 = 7,$	$\mathcal{PL} = \{1\},$	$\mathcal{TL} = \{2, 3, 4\}$
2.	$L_3 = \min\{\tilde{L}_2, \tilde{L}_3, \tilde{L}_4\} = 5, k = 3,$	$\mathcal{PL} = \{1, 3\},$	$\mathcal{TL} = \{2, 4\}$
3.	$\tilde{L}_2 = \min\{8, L_3 + l_{32}\} = \min\{8, 5 + 1\} = 6$		
	$\tilde{L}_4 = \min\{7, L_3 + l_{34}\} = \min\{7, \infty\} = 7$		
2.	$L_2 = \min\{\tilde{L}_2, \tilde{L}_4\} = \min\{6, 7\} = 6, k = 2,$	$\mathcal{PL} = \{1, 2, 3\},$	$\mathcal{TL} = \{4\}$
3.	$\tilde{L}_4 = \min\{7, L_2 + l_{24}\} = \min\{7, 6 + 2\} = 7$		
2.	$L_4 = 7, k = 4$	$\mathcal{PL} = \{1, 2, 3, 4\},$	$\mathcal{TL} = \varnothing.$

Figure 487b shows the resulting shortest paths, of lengths $L_2 = 6$, $L_3 = 5$, $L_4 = 7$. ■



(a) Given graph G (b) Shortest paths in G

Fig. 487. Example 1

Complexity. *Dijkstra's algorithm is* $O(n^2)$.

PROOF Step 2 requires comparison of elements, first $n - 2$, the next time $n - 3$, etc., a total of $(n - 2)(n - 1)/2$. Step 3 requires the same number of comparisons, a total of $(n - 2)(n - 1)/2$, as well as additions, first $n - 2$, the next time $n - 3$, etc., again a total of $(n - 2)(n - \ )/2$. Hence the total number of operations is $3(n - 2)(n - 1)/2 = O(n^2)$. ■

PROBLEM SET 23.3

1. The net of roads in Fig. 488 connecting four villages is to be reduced to minimum length, but so that one can still reach every village from every other village. Which of the roads should be retained? Find the solution **(a)** by inspection, **(b)** by Dijkstra's algorithm.

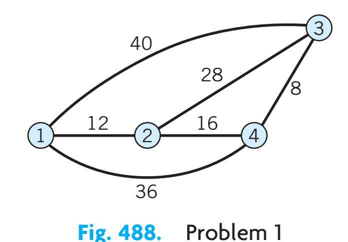


Fig. 488. Problem 1

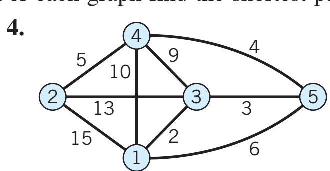
2. Show that in Dijkstra's algorithm, for L_k there is a path $P: 1 \rightarrow k$ of length L_k.

3. Show that in Dijkstra's algorithm, at each instant the demand on storage is light (data for fewer than n edges).

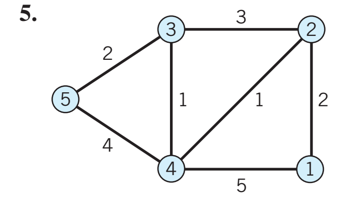
4–9 DIJKSTRA'S ALGORITHM

For each graph find the shortest paths.

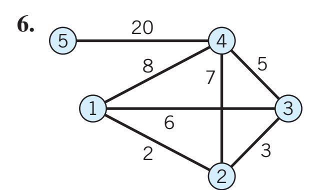
4.



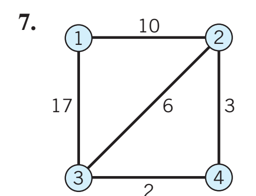
5.



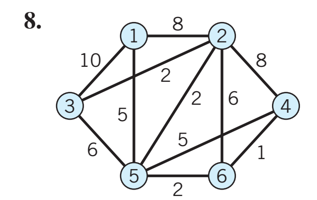
6.



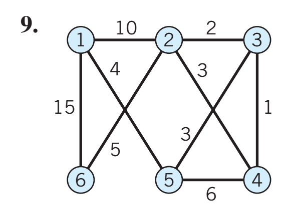
7.



8.



9.



23.4 Shortest Spanning Trees: Greedy Algorithm

So far we have discussed shortest path problems. We now turn to a particularly important kind of graph, called a tree, along with related optimization problems that arise quite often in practice.

By definition, a **tree** T is a graph that is connected and has no cycles. "**Connected**" was defined in Sec. 23.3; it means that there is a path from any vertex in T to any other vertex in T. A **cycle** is a path $s \rightarrow t$ of at least three edges that is closed $(t = s)$; see also Sec. 23.2. Figure 489a shows an example.

CAUTION! The terminology varies; *cycles* are sometimes also called *circuits.*

A **spanning tree** T in a given connected graph $G = (V, E)$ is a tree containing *all* the n vertices of G. See Fig. 489b. Such a tree has $n - 1$ edges. (Proof?)

A **shortest spanning tree** T in a connected graph G (whose edges (i, j) have lengths $l_{ij} > 0$) is a spanning tree for which Σl_{ij} (sum over all edges of T) is minimum compared to Σl_{ij} for any other spanning tree in G.

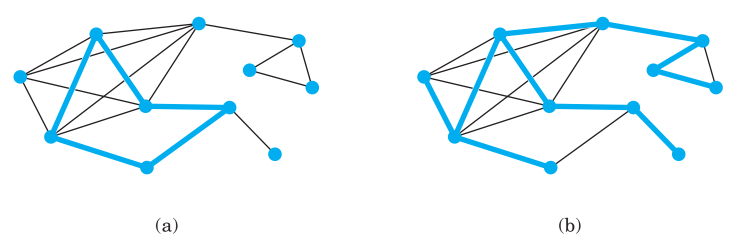


Fig. 489. Example of (a) a cycle, (b) a spanning tree in a graph

Trees are among the most important types of graphs, and they occur in various applications. Familiar examples are family trees and organization charts. Trees can be used to exhibit, organize, or analyze electrical networks, producer–consumer and other business relations, information in database systems, syntactic structure of computer programs, etc. We mention a few specific applications that need no lengthy additional explanations.

The set of shortest paths from vertex 1 to the vertices $2, \cdots, n$ in the last section forms a spanning tree.

Railway lines connecting a number of cities (the vertices) can be set up in the form of a spanning tree, the "length" of a line (edge) being the construction cost, and one wants to minimize the total construction cost. Similarly for bus lines, where "length" may be

the average annual operating cost. Or for steamship lines (freight lines), where "length" may be profit and the goal is the maximization of total profit. Or in a network of telephone lines between some cities, a shortest spanning tree may simply represent a selection of lines that connect all the cities at minimal cost. In addition to these examples we could mention others from distribution networks, and so on.

We shall now discuss a simple algorithm for the problem of finding a shortest spanning tree. This algorithm (Table 23.3) is particularly suitable for sparse graphs (graphs with very few edges; see Sec. 23.1).

Table 23.3 Kruskal's[5] Greedy Algorithm for Shortest Spanning Trees

Proceedings of the American Mathematical Society **7** (1956), 48–50.

ALGORITHM KRUSKAL $[G = (V, E), l_{ij}$ for all (i, j) in $E]$

Given a connected graph $G = (V, E)$ with vertices $1, 2, \cdots, n$ and edges (i, j) having length $l_{ij} > 0$, the algorithm determines a shortest spanning tree T in G.

INPUT: Edges (i, j) of G and their lengths l_{ij}

OUTPUT: Shortest spanning tree T in G

1. Order the edges of G in ascending order of length.

2. Choose them in this order as edges of $\boldsymbol{T}$, rejecting an edge only if it forms a cycle with edges already chosen.

If $n - 1$ edges have been chosen, then

OUTPUT T (= the set of edges chosen). Stop

End KRUSKAL

EXAMPLE 1 Application of Kruskal's Algorithm

Using Kruskal's algorithm, we shall determine a shortest spanning tree in the graph in Fig. 490.

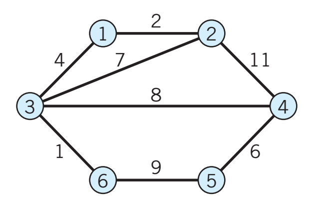


Fig. 490. Graph in Example 1

Table 23.4 Solution in Example 1

Edge	Length	Choice
(3, 6)	1	1st
(1, 2)	2	2nd
(1, 3)	4	3rd
(4, 5)	6	4th
(2, 3)	7	Reject
(3, 4)	8	5th
(5, 6)	9	
(2, 4)	11	

Solution. See Table 23.4. In some of the intermediate stages the edges chosen form a *disconnected* graph (see Fig. 491); this is typical. We stop after $n - 1 = 5$ choices since a spanning tree has $n - 1$ edges. In our problem the edges chosen are in the upper part of the list. This is typical of problems of any size; in general, edges farther down in the list have a smaller chance of being chosen. ■

[5]JOSEPH BERNARD KRUSKAL (1928–), American mathematician who worked at Bell Laboratories. He is known for his contributions to graph theory and statistics.

The efficiency of Kruskal's method is greatly increased by double labeling of vertices.

Double Labeling of Vertices. *Each vertex i carries a double label* (r_i, p_i), *where*

$$r_i = \textit{Root of the subtree to which } i \textit{ belongs,}$$

$$p_i = \textit{Predecessor of } i \textit{ in its subtree,}$$

$$p_i = 0 \textit{ for roots.}$$

This simplifies rejecting.

Rejecting. If (i, j) *is next in the list to be considered, reject* (i, j) *if* $r_i = r_j$ (that is, i and j are in the same subtree, so that they are already joined by edges and (i, j) would thus create a cycle). *If* $r_i \neq r_j$, *include* (i, j) *in T.*

If there are several choices for r_i, choose the smallest. If subtrees merge (become a single tree), retain the smallest root as the root of the new subtree.

For Example 1 the double-label list is shown in Table 23.5. In storing it, at each instant one may retain only the latest double label. We show all double labels in order to exhibit the process in all its stages. Labels that remain unchanged are not listed again. Underscored are the two 1's that are the common root of vertices 2 and 3, the reason for rejecting the edge (2, 3). By reading for each vertex the latest label we can read from this list that 1 is the vertex we have chosen as a root and the tree is as shown in the last part of Fig. 491.

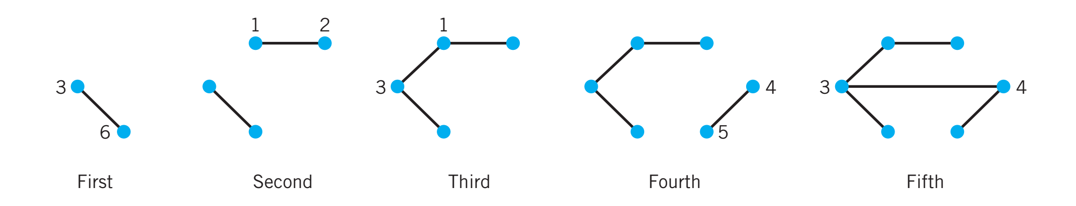


Fig. 491. Choice process in Example 1

Table 23.5 List of Double Labels in Example 1

Vertex	Choice 1 (3, 6)	Choice 2 (1, 2)	Choice 3 (1, 3)	Choice 4 (4, 5)	Choice 5 (3, 4)
1		(1, 0)			
2		($\underline{1}$, 1)			
3	(3, 0)		($\underline{1}$, 1)		
4				(4, 0)	(1, 3)
5				(4, 4)	(1, 4)
6	(3, 3)		(1, 3)		

This is made possible by the predecessor label that each vertex carries. Also, for accepting or rejecting an edge we have to make only one comparison (the roots of the two endpoints of the edge).

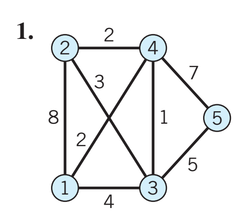
Ordering is the more expensive part of the algorithm. It is a standard process in data processing for which various methods have been suggested (see **Sorting** in Ref. [E25] listed in App. 1). For a complete list of m edges, an algorithm would be $O(m \log_2 m)$, but since the $n - 1$ edges of the tree are most likely to be found earlier, by inspecting the q $(< m)$ topmost edges, for such a list of q edges one would have $O(q \log_2 m)$.

PROBLEM SET 23.4

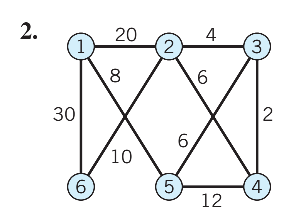
1–6 KRUSKAL'S GREEDY ALGORITHM

Find a shortest spanning tree by Kruskal's algorithm. Sketch it.

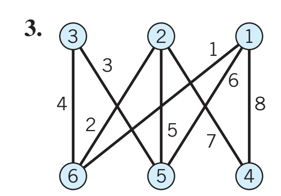
1.



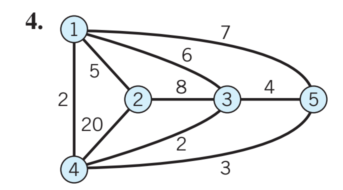
2.



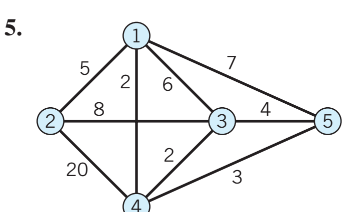
3.



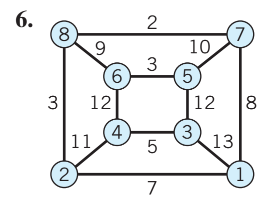
4.



5.



6.



7. CAS PROBLEM. Kruskal's Algorithm. Write a corresponding program. (Sorting is discussed in Ref. [E25] listed in App. 1.)

8. To get a minimum spanning tree, instead of adding shortest edges, one could think of deleting longest edges. For what graphs would this be feasible? Describe an algorithm for this.

9. Apply the method suggested in Prob. 8 to the graph in Example 1. Do you get the same tree?

10. Design an algorithm for obtaining longest spanning trees.

11. Apply the algorithm in Prob. 10 to the graph in Example 1. Compare with the result in Example 1.

12. Forest. A (not necessarily connected) graph without cycles is called a *forest*. Give typical examples of applications in which graphs occur that are forests or trees.

	Dallas	Denver	Los Angeles	New York	Washington, DC
Chicago	800	900	1800	700	650
Dallas		650	1300	1350	1200
Denver			850	1650	1500
Los Angeles				2500	2350
New York					200

13. Air cargo. Find a shortest spanning tree in the complete graph of all possible 15 connections between the six cities given (distances by airplane, in miles, rounded). Can you think of a practical application of the result?

14–20 GENERAL PROPERTIES OF TREES

Prove the following. *Hint.* Use Prob. 14 in proving 15 and 18; use Probs. 16 and 18 in proving 20.

14. Uniqueness. The path connecting any two vertices u and v in a tree is unique.

15. If in a graph any two vertices are connected by a unique path, the graph is a tree.

16. If a graph has no cycles, it must have at least 2 vertices of degree 1 (definition in Sec. 23.1).

17. A tree with exactly two vertices of degree 1 must be a path.

18. A tree with n vertices has $n - 1$ edges. (Proof by induction.)

19. If two vertices in a tree are joined by a new edge, a cycle is formed.

20. A graph with n vertices is a tree if and only if it has $n - 1$ edges and has no cycles.

23.5 Shortest Spanning Trees: Prim's Algorithm

Prim's[6] algorithm, shown in Table 23.6, is another popular algorithm for the shortest spanning tree problem (see Sec. 23.4). This algorithm avoids ordering edges and gives a tree T at each stage, a property that Kruskal's algorithm in the last section did not have (look back at Fig. 491 if you did not notice it).

In Prim's algorithm, starting from any single vertex, which we call 1, we "grow" the tree T by adding edges to it, one at a time, according to some rule (in Table 23.6) until T finally becomes a *spanning* tree, which is shortest.

We denote by U the set of vertices of the growing tree T and by S the set of its edges. Thus, initially $U = \{1\}$ and $S = \emptyset$; at the end, $U = V$, the vertex set of the given graph $G = (V, E)$, whose edges (i, j) have length $l_{ij} > 0$, as before.

[6]ROBERT CLAY PRIM (1921–), American computer scientist at General Electric, Bell Laboratories, and Sandia National Laboratories.

Thus at the beginning (Step 1) the labels

$$\lambda_2, \cdots, \lambda_n \qquad \text{of the vertices} \qquad 2, \cdots, n$$

are the lengths of the edges connecting them to vertex 1 (or ∞ if there is no such edge in G). And we pick (Step 2) the shortest of these as the first edge of the growing tree T and include its other end j in U (choosing the smallest j if there are several, to make the process unique). Updating labels in Step 3 (at this stage and at any later stage) concerns each vertex k not yet in U. Vertex k has label $\lambda_k = l_{i(k),k}$ from before. If $l_{jk} < \lambda_k$, this means that k is closer to the new member j just included in U than k is to its old "closest neighbor" $i(k)$ in U. Then we update the label of k, replacing $\lambda_k = l_{i(k),k}$ by $\lambda_k = l_{jk}$ and setting $i(k) = j$. If, however, $l_{jk} \geqq \lambda_k$ (the *old* label of k), we don't touch the old label. Thus the label λ_k always identifies the closest neighbor of k in U, and this is updated in Step 3 as U and the tree T grow. From the final labels we can backtrack the final tree, and from their numeric values we compute the total length (sum of the lengths of the edges) of this tree.

Prim's algorithm is useful for computer network design, cable, distribution networks, and transportation networks.

Table 23.6 Prim's Algorithm for Shortest Spanning Trees

Bell System Technical Journal **36** (1957), 1389–1401.

For an improved version of the algorithm, see Cheriton and Tarjan, *SIAM Journal on Computation* **5** (1976), 724–742.

ALGORITHM PRIM [$G = (V, E)$, $V = \{1, \cdots, n\}$, l_{ij} for all (i, j) in E]

Given a connected graph $G = (V, E)$ with vertices $1, 2, \cdots, n$ and edges (i, j) having length $l_{ij} > 0$, this algorithm determines a shortest spanning tree T in G and its length $L(T)$.

INPUT: n, edges (i, j) of G and their lengths l_{ij}
OUTPUT: Edge set S of a shortest spanning tree T in G; $L(T)$
[*Initially, all vertices are unlabeled.*]

1. *Initial step*
Set $i(k) = 1$, $U = \{1\}$, $S = \emptyset$.
Label vertex k $(= 2, \cdots, n)$ with $\lambda_k = l_{ik}$ [$= \infty$ if G has no edge $(1, k)$].

2. *Addition of an edge to the tree T*
Let λ_j be the smallest λ_k for vertex k not in U. Include vertex j in U and edge $(i(j), j)$ in S.
If $U = V$ then compute
$L(T) = \Sigma l_{ij}$ (sum over all edges in S)
OUTPUT S, $L(T)$. Stop
[*S is the edge set of a shortest spanning tree T in G.*]
Else continue (that is, go to Step 3).

3. *Label updating*
For every k not in U, if $l_{jk} < \lambda_k$, then set $\lambda_k = l_{jk}$ and $i(k) = j$.
Go to Step 2.

End PRIM

EXAMPLE 1 Application of Prim's Algorithm

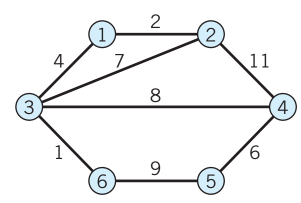


Fig. 492. Graph in Example 1

Find a shortest spanning tree in the graph in Fig. 492 (which is the same as in Example 1, Sec. 23.4, so that we can compare).

Solution. The steps are as follows.

1. $i(k) = 1$, $U = \{1\}$, $S = \varnothing$, initial labels see Table 23.7.
2. $\lambda_2 = l_{12} = 2$ is smallest, $U = \{1, 2\}$, $S = \{(1, 2)\}$.
3. Update labels as shown in Table 23.7, column (I).
2. $\lambda_3 = l_{13} = 4$ is smallest, $U = \{1, 2, 3\}$, $S = \{(1, 2), (1, 3)\}$.
3. Update labels as shown in Table 23.7, column (II).
2. $\lambda_6 = l_{36} = 1$ is smallest, $U = \{1, 2, 3, 6\}$, $S = \{(1, 2), (1, 3), (3, 6)\}$.
3. Update labels as shown in Table 23.7, column (III).
2. $\lambda_4 = l_{34} = 8$ is smallest, $U = \{1, 2, 3, 4, 6\}$, $S = \{(1, 2), (1, 3), (3, 4), (3, 6)\}$.
3. Update labels as shown in Table 23.7, column (IV).
2. $\lambda_5 = l_{45} = 6$ is smallest, $U = V$, $S = (1, 2), (1, 3), (3, 4), (3, 6), (4, 5)$. Stop.

The tree is the same as in Example 1, Sec. 23.4. Its length is 21. You will find it interesting to compare the growth process of the present tree with that in Sec. 23.4. ■

Table 23.7 **Labeling of Vertices in Example 1**

Vertex	Initial Label	Relabeling			
		(I)	(II)	(III)	(IV)
2	$l_{12} = 2$	—	—	—	—
3	$l_{13} = 4$	$l_{13} = 4$	—	—	—
4	∞	$l_{24} = 11$	$l_{34} = 8$	$l_{34} = 8$	—
5	∞	∞	∞	$l_{65} = 9$	$l_{45} = 6$
6	∞	∞	$l_{36} = 1$	—	—

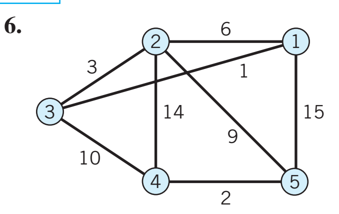
PROBLEM SET 23.5

SHORTEST SPANNING TREES. PRIM'S ALGORITHM

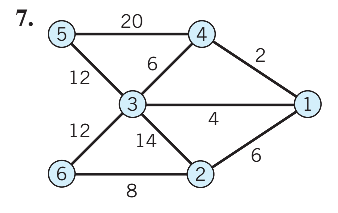
1. When will $S = E$ at the end in Prim's algorithm?
2. **Complexity.** Show that Prim's algorithm has complexity $O(n^2)$.
3. What is the result of applying Prim's algorithm to a graph that is not connected?
4. If for a complete graph (or one with very few edges missing), our data is an $n \times n$ distance table (as in Prob. 13, Sec. 23.4), show that the present algorithm [which is $O(n^2)$] cannot easily be replaced by an algorithm of order less than $O(n^2)$.
5. How does Prim's algorithm prevent the generation of cycles as you grow T?

6–13 Find a shortest spanning tree by Prim's algorithm.

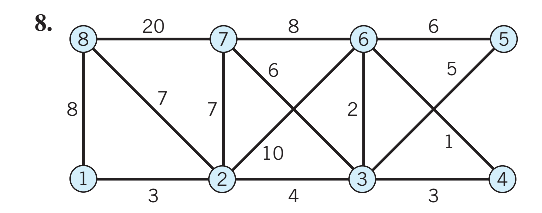
6.



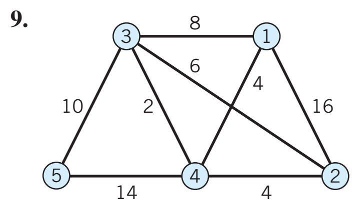
7.



8.



9.



10. For the graph in Prob. 6, Sec. 23.4.

11. For the graph in Prob. 4, Sec. 23.4.

12. For the graph in Prob. 2, Sec. 23.4.

13. CAS PROBLEM. Prim's Algorithm. Write a program and apply it to Probs. 6–9.

14. TEAM PROJECT. Center of a Graph and Related Concepts. (a) Distance, Eccentricity. Call the length of a shortest path $u \to v$ in a graph $G = (V, E)$ the *distance* $d(u, v)$ from u to v. For fixed u, call the greatest $d(u, v)$ as v ranges over V the *eccentricity* $\epsilon(u)$ of u. Find the eccentricity of vertices 1, 2, 3 in the graph in Prob. 7.

(b) Diameter, Radius, Center. The *diameter* $d(G)$ of a graph $G = (V, E)$ is the maximum of $d(u, v)$ as u and v vary over V, and the *radius* $r(G)$ is the smallest eccentricity $\epsilon(v)$ of the vertices v. A vertex v with $\epsilon(v) = r(G)$ is called a *central vertex*. The set of all central vertices is called the *center* of G. Find $d(G)$, $r(G)$, and the center of the graph in Prob. 7.

(c) What are the diameter, radius, and center of the spanning tree in Example 1 of the text?

(d) Explain how the idea of a center can be used in setting up an emergency service facility on a transportation network. In setting up a fire station, a shopping center. How would you generalize the concepts in the case of two or more such facilities?

(e) Show that a tree T whose edges all have length 1 has center consisting of either one vertex or two adjacent vertices.

(f) Set up an algorithm of complexity $O(n)$ for finding the center of a tree T.

23.6 Flows in Networks

After shortest path problems and problems for trees, as a third large area in combinatorial optimization we discuss **flow problems in networks** (electrical, water, communication, traffic, business connections, etc.), turning from graphs to digraphs (directed graphs; see Sec. 23.1).

By definition, a **network** is a digraph $G = (V, E)$ in which each edge (i, j) has assigned to it a **capacity** $c_{ij} > 0$ [= maximum possible flow along (i, j)], and at one vertex, s, called the **source**, a flow is produced that flows along the edges of the digraph G to another vertex, t, called the **target** or **sink**, where the flow disappears.

In applications, this may be the flow of electricity in wires, of water in pipes, of cars on roads, of people in a public transportation system, of goods from a producer to consumers, of e-mail from senders to recipients over the Internet, and so on.

We denote the flow along a (directed!) edge (i, j) by f_{ij} and impose two conditions:

1. For each edge (i, j) in G the flow does not exceed the capacity c_{ij},

$$0 \leqq f_{ij} \leqq c_{ij} \qquad \text{(“Edge condition”).} \tag{1}$$

2. For each vertex i, not s or t,

$$\text{Inflow} = \text{Outflow} \qquad \text{(“Vertex condition,” “Kirchhoff's law”);}$$

in a formula,

$$(2)\qquad \underbrace{\sum_k f_{ki}}_{\text{Inflow}} - \underbrace{\sum_j f_{ij}}_{\text{Outflow}} = \begin{cases} 0 \text{ if vertex } i \neq s,\ i \neq t, \\ -f \text{ at the source } s, \\ f \text{ at the target (sink) } t, \end{cases}$$

where f is the total flow (and at s the inflow is zero, whereas at t the outflow is zero). Figure 493 illustrates the notation (for some hypothetical figures).

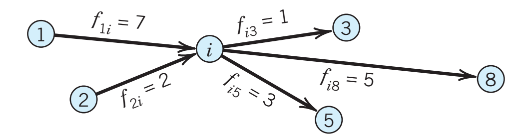


Fig. 493. Notation in (2): inflow and outflow for a vertex i (not s or t)

Paths

By a **path** $v_1 \to v_k$ from a vertex v_1 to a vertex v_k in a digraph G we mean a sequence of edges

$$(v_1, v_2), (v_2, v_3), \cdots, (v_{k-1}, v_k),$$

regardless of their directions in G, that forms a path as in a graph (see Sec. 23.2). Hence when we travel along this path from v_1 to v_k we may traverse some edge ***in*** its given direction—then we call it a **forward edge** of our path—or ***opposite to*** its given direction—then we call it a **backward edge** of our path. In other words, our path consists of one-way streets, and forward edges (backward edges) are those that we travel *in the right direction* (*in the wrong direction*). Figure 494 shows a forward edge (u, v) and a backward edge (w, v) of a path $v_1 \to v_k$.

CAUTION! Each edge in a network has a given direction, *which we cannot change*. Accordingly, if (u, v) is a forward edge in a path $v_1 \to v_k$, then (u, v) can become a backward edge only in another path $x_1 \to x_j$ in which it is an edge and is traversed in the opposite direction as one goes from x_1 to x_j; see Fig. 495. Keep this in mind, to avoid misunderstandings.

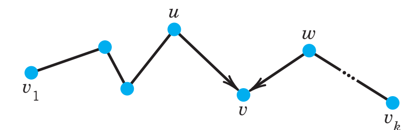


Fig. 494. Forward edge (u, v) and backward edge (w, v) of a path $v_1 \to v_k$

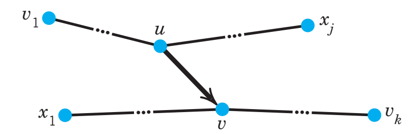


Fig. 495. Edge (u, v) as forward edge in the path $v_1 \to v_k$ and as backward edge in the path $x_1 \to x_j$

Flow Augmenting Paths

Our goal will be to ***maximize the flow*** from the source s to the target t of a given network. We shall do this by developing methods for increasing an existing flow (including the special case in which the latter is zero). The idea then is to find a path $P: s \to t$ all of

whose edges are not fully used, so that we can push additional flow through P. This suggests the following concept.

DEFINITION

Flow Augmenting Path

A *flow augmenting path* in a network with a given flow f_{ij} on each edge (i, j) is a path $P: s \to t$ such that

(i) no forward edge is used to capacity; thus $f_{ij} < c_{ij}$ for these;

(ii) no backward edge has flow 0; thus $f_{ij} > 0$ for these.

EXAMPLE 1 Flow Augmenting Paths

Find flow augmenting paths in the network in Fig. 496, where the first number is the capacity and the second number a given flow.

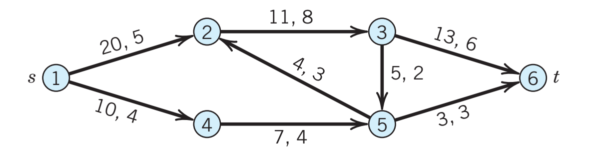


Fig. 496. Network in Example 1
First number = Capacity, Second number = Given flow

Solution. In practical problems, networks are large and one needs a ***systematic method for augmenting flows, which we discuss in the next section.*** In our small network, which should help to illustrate and clarify the concepts and ideas, we can find flow augmenting paths by inspection and augment the existing flow $f = 9$ in Fig. 496. (The outflow from s is $5 + 4 = 9$, which equals the inflow $6 + 3$ into t.)

We use the notation

$$\Delta_{ij} = c_{ij} - f_{ij} \qquad \text{for forward edges}$$

$$\Delta_{ij} = f_{ij} \qquad \text{for backward edges}$$

$$\Delta = \min \Delta_{ij} \qquad \text{taken over all edges of a path.}$$

From Fig. 496 we see that a flow augmenting path $P_1: s \to t$ is $P_1: 1 - 2 - 3 - 6$ (Fig. 497), with $\Delta_{12} = 20 - 5 = 15$, etc., and $\Delta = 3$. Hence we can use P_1 to increase the given flow 9 to $f = 9 + 3 = 12$. All three edges of P_1 are forward edges. We augment the flow by 3. Then the flow in each of the edges of P_1 is increased by 3, so that we now have $f_{12} = 8$ (instead of 5), $f_{23} = 11$ (instead of 8), and $f_{36} = 9$ (instead of 6). Edge (2, 3) is now used to capacity. The flow in the other edges remains as before.

We shall now try to increase the flow in this network in Fig. 496 beyond $f = 12$.

There is another flow augmenting path $P_2: s \to t$, namely, $P_2: 1 - 4 - 5 - 3 - 6$ (Fig. 497). It shows how a backward edge comes in and how it is handled. Edge (3, 5) is a backward edge. It has flow 2, so that $\Delta_{36} = 2$. We compute $\Delta_{14} = 10 - 4 = 6$, etc. (Fig. 497) and $\Delta = 2$. Hence we can use P_2 for another augmentation to get $f = 12 + 2 = 14$. The new flow is shown in Fig. 498. No further augmentation is possible. We shall confirm later that $f = 14$ is maximum. ■

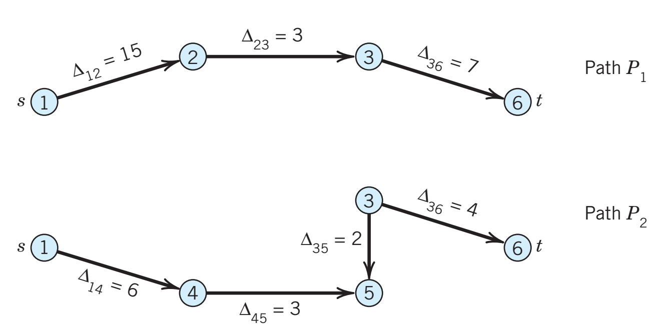


Fig. 497. Flow augmenting paths in Example 1

Cut Sets

A **cut set** is a set of edges in a network. The underlying idea is simple and natural. If we want to find out what is flowing from s to t in a network, we may cut the network somewhere between s and t (Fig. 498 shows an example) and see what is flowing in the edges hit by the cut, because any flow from s to t must sometimes pass through some of these edges. These form what is called a **cut set**. [In Fig. 498, the cut set consists of the edges (2, 3), (5, 2), (4, 5).] We denote this cut set by (S, T). Here S is the set of vertices on that side of the cut on which s lies ($S = \{s, 2, 4\}$ for the cut in Fig. 498) and T is the set of the other vertices ($T = \{3, 5, t\}$ in Fig. 498). We say that a cut ***partitions*** the vertex set V into two parts S and T. Obviously, the corresponding cut set (S, T) consists of all the edges in the network with one end in S and the other end in T.

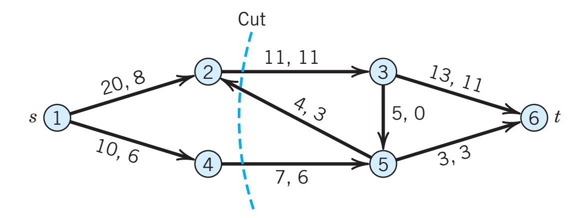


Fig. 498. Maximum flow in Example 1

By definition, the **capacity** cap (S, T) of a cut set (S, T) is the sum of the capacities of all **forward edges** in (S, T) (forward edges only!), that is, the edges that are directed *from S to T*,

(3) $$\text{cap}\,(S, T) = \Sigma c_{ij} \qquad \text{[sum over the forward edges of } (S, T)].$$

Thus, cap $(S, T) = 11 + 7 = 18$ in Fig. 498.

Explanation. This can be seen as follows. Look at Fig. 498. Recall that for each edge in that figure, the first number denotes capacity and the second number flow. Intuitively, you can think of the edges as roads, where the capacity of the road is how many cars can actually be on the road, and the flow denotes how many cars actually are on the road. To compute capacity cap (S, T) we are only looking at the first number on the edges. Take a look and see that the cut physically cuts three edges, that is, (2, 3), (4, 5), and (5, 2). *The cut concerns only forward edges* that are being cut, so it concerns edges (2, 3) and (4, 5) (and does not include edge (5, 2) which is also being cut, but since it goes backwards, it does not count). Hence (2, 3) contributes 11 and (4, 5) contributes 7 to the capacity cap (S, T), for a total of 18 in Fig. 498. Hence cap $(S, T) = 18$.

The other edges (directed *from T to S*) are called **backward edges** of the cut set (S, T), and by the **net flow** through a cut set we mean the sum of the flows in the forward edges minus the sum of the flows in the backward edges of the cut set.

CAUTION! Distinguish well between forward and backward edges in a cut set and in a path: (5, 2) in Fig. 498 is a backward edge for the cut shown but a forward edge in the path $1 - 4 - 5 - 2 - 3 - 6$.

For the cut in Fig. 498 the net flow is $11 + 6 - 3 = 14$. For the same cut in Fig. 496 (not indicated there), the net flow is $8 + 4 - 3 = 9$. In both cases it equals the flow f.

We claim that this is not just by chance, but cuts do serve the purpose for which we have introduced them:

THEOREM 1

Net Flow in Cut Sets

Any given flow in a network G is the net flow through any cut set (S, T) of G.

PROOF By Kirchhoff's law (2), multiplied by -1, at a vertex i we have

$$(4) \qquad \underbrace{\sum_j f_{ij}}_{\text{Outflow}} - \underbrace{\sum_l f_{li}}_{\text{Inflow}} = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

Here we can sum over j and l from 1 to n (= number of vertices) by putting $f_{ij} = 0$ for $j = i$ and also for edges without flow or nonexisting edges; hence we can write the two sums as one,

$$\sum_j (f_{ij} - f_{ji}) = \begin{cases} 0 & \text{if } i \neq s, t, \\ f & \text{if } i = s. \end{cases}$$

We now sum over all i in S. Since s is in S, this sum equals f:

$$(5) \qquad \sum_{i \in S} \sum_{j \in V} (f_{ij} - f_{ji}) = f.$$

We claim that in this sum, only the edges belonging to the cut set contribute. Indeed, edges with both ends in T cannot contribute, since we sum only over i in S; but edges (i, j) with both ends in S contribute $+f_{ij}$ at one end and $-f_{ij}$ at the other, a total contribution of 0. Hence the left side of (5) equals the net flow through the cut set. By (5), this is equal to the flow f and proves the theorem. ∎

This theorem has the following consequence, which we shall also need later in this section.

THEOREM 2

Upper Bound for Flows

A flow f in a network G cannot exceed the capacity of any cut set (S, T) in G.

PROOF By Theorem 1 the flow f equals the net flow through the cut set, $f = f_1 - f_2$, where f_1 is the sum of the flows through the forward edges and f_2 ($\geqq 0$) is the sum of the flows through the backward edges of the cut set. Thus $f \leqq f_1$. Now f_1 cannot exceed the sum of the capacities of the forward edges; but this sum equals the capacity of the cut set, by definition. Together, $f \leqq \text{cap}\,(S, T)$, as asserted. ∎

Cut sets will now bring out the full importance of augmenting paths:

THEOREM 3

Main Theorem. Augmenting Path Theorem for Flows

A flow from s to t in a network G is maximum if and only if there does not exist a flow augmenting path $s \to t$ in G.

PROOF **(a)** If there is a flow augmenting path $P: s \to t$, we can use it to push through it an additional flow. Hence the given flow cannot be maximum.

(b) On the other hand, suppose that there is no flow augmenting path $s \to t$ in G. Let S_0 be the set of all vertices i (including s) such that there is a flow augmenting path $s \to i$, and let T_0 be the set of the other vertices in G. Consider any edge (i, j) with i in S_0 and j in T_0. Then we have a flow augmenting path $s \to i$ since i is in S_0, but $s \to i \to j$ is not flow augmenting because j is not in S_0. Hence we must have

$$(6) \qquad f_{ij} = \begin{cases} c_{ij} \\ 0 \end{cases} \quad \text{if } (i, j) \text{ is a } \begin{cases} \text{forward} \\ \text{backward} \end{cases} \text{edge of the path } s \to i \to j.$$

Otherwise we could use (i, j) to get a flow augmenting path $s \to i \to j$. Now (S_0, T_0) defines a cut set (since t is in T_0; why?). Since by (6), forward edges are used to capacity and backward edges carry no flow, the net flow through the cut set (S_0, T_0) equals the sum of the capacities of the forward edges, which is cap (S_0, T_0) by definition. This net flow equals the given flow f by Theorem 1. Thus $f =$ cap (S_0, T_0). We also have $f \leqq$ cap (S_0, T_0) by Theorem 2. Hence f must be maximum since we have reached equality. ■

The end of this proof yields another basic result (by Ford and Fulkerson, *Canadian Journal of Mathematics* **8** (1956), 399–404), namely, the so-called

THEOREM 4

Max-Flow Min-Cut Theorem

The maximum flow in any network G equals the capacity of a "**minimum cut set**" (= a cut set of minimum capacity) *in G.*

PROOF We have just seen that $f =$ cap (S_0, T_0) for a maximum flow f and a suitable cut set (S_0, T_0). Now by Theorem 2 we also have $f \leqq$ cap (S, T) for this f and any cut set (S, T) in G. Together, cap $(S_0, T_0) \leqq$ cap (S, T). Hence (S_0, T_0) is a minimum cut set.

The existence of a maximum flow in this theorem follows for rational capacities from the algorithm in the next section and for arbitrary capacities from the Edmonds–Karp BFS also in that section. ■

The two basic tools in connection with networks are flow augmenting paths and cut sets. In the next section we show how flow augmenting paths can be used in an algorithm for maximum flows.

PROBLEM SET 23.6

1–6 CUT SETS, CAPACITY

Find T and cap (S, T) for:

1. Fig. 498, $S = \{1, 2, 4, 5\}$
2. Fig. 499, $S = \{1, 2, 3\}$
3. Fig. 498, $S = \{1, 2, 3\}$
4. Fig. 499, $S = \{1, 2\}$
5. Fig. 499, $S = \{1, 2, 4, 5\}$
6. Fig. 498, $S = \{1, 3, 5\}$

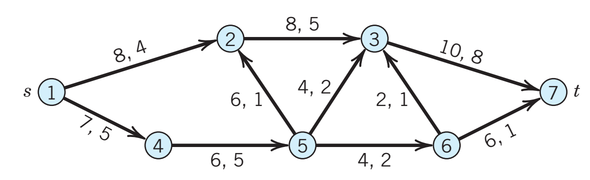


Fig. 499. Problems 2, 4, and 5

7–8 MINIMUM CUT SET

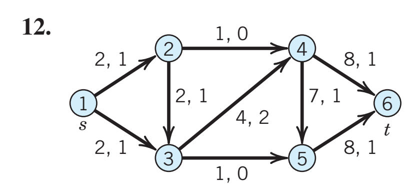
Find a minimum cut set and its capacity for the network:

7. In Fig. 499
8. In Fig. 496. Verify that its capacity equals the maximum flow.
9. Why are backward edges not considered in the definition of the capacity of a cut set?
10. **Incremental network.** Sketch the network in Fig. 499, and on each edge (i, j) write $c_{ij} - f_{ij}$ and f_{ij}. Do you recognize that from this "incremental network" one can more easily see flow augmenting paths?
11. **Omission of edges.** Which edges could be omitted from the network in Fig. 499 without decreasing the maximum flow?

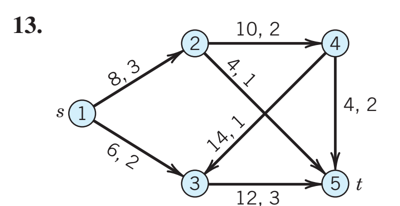
12–15 FLOW AUGMENTING PATHS

Find flow augmenting paths:

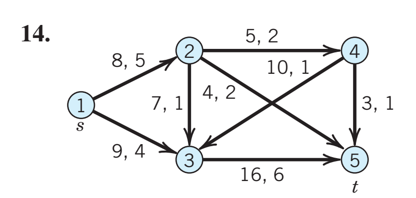
12.



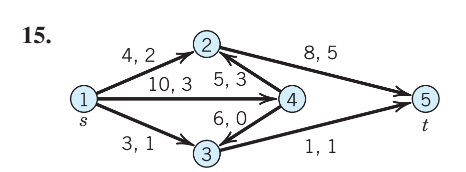
13.



14.



15.

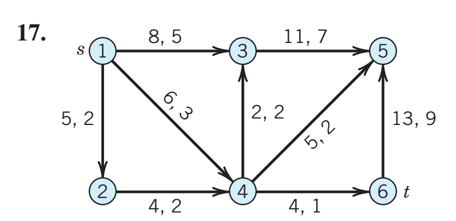


16–19 MAXIMUM FLOW

Find the maximum flow by inspection:

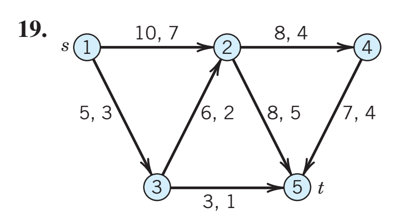
16. In Prob. 13

17.



18. In Prob. 12

19.



20. Find another maximum flow $f = 15$ in Prob. 19.

23.7 Maximum Flow: Ford–Fulkerson Algorithm

Flow augmenting paths, as discussed in the last section, are used as the basic tool in the Ford–Fulkerson[7] algorithm in Table 23.8 in which a given flow (for instance, zero flow in all edges) is increased until it is maximum. The algorithm accomplishes the increase by a stepwise construction of flow augmenting paths, one at a time, until no further such paths can be constructed, which happens precisely when the flow is maximum.

In Step 1, an initial flow may be given. In Step 3, a vertex j can be labeled if there is an edge (i, j) with i labeled and

$$c_{ij} > f_{ij} \qquad \textbf{(“forward edge”)}$$

or if there is an edge (j, i) with i labeled and

$$f_{ji} > 0 \qquad \textbf{(“backward edge”)}.$$

To **scan** a labeled vertex i means to label every unlabeled vertex j adjacent to i that can be labeled. Before scanning a labeled vertex i, scan all the vertices that got labeled before i. This **BFS** (**Breadth First Search**) strategy was suggested by Edmonds and Karp in 1972 (*Journal of the Association for Computing Machinery* **19**, 248–64). It has the effect that one gets shortest possible augmenting paths.

Table 23.8 Ford–Fulkerson Algorithm for Maximum Flow

Canadian Journal of Mathematics **9** (1957), 210–218

ALGORITHM FORD–FULKERSON

$[G = (V, E)$, vertices $1\ (= s), \cdots, n\ (= t)$, edges (i, j), $c_{ij}]$

This algorithm computes the maximum flow in a network G with source s, sink t, and capacities $c_{ij} > 0$ of the edges (i, j).

INPUT: n, $s = 1$, $t = n$, edges (i, j) of G, c_{ij}

OUTPUT: Maximum flow f in G

1. Assign an initial flow f_{ij} (for instance, $f_{ij} = 0$ for all edges), compute f.
2. Label s by $\varnothing$. Mark the other vertices *“unlabeled.”*
3. Find a labeled vertex i that has not yet been scanned. Scan i as follows. For every unlabeled adjacent vertex j, if $c_{ij} > f_{ij}$, compute

$$\Delta_{ij} = c_{ij} - f_{ij} \quad \text{and} \quad \Delta_j = \begin{cases} \Delta_{ij} & \text{if } i = 1 \\ \min(\Delta_i, \Delta_{ij}) & \text{if } i > 1 \end{cases}$$

and label j with a *“forward label”* (i^+, Δ_j); or if $f_{ji} > 0$, compute

$$\Delta_j = \min(\Delta_i, f_{ji})$$

and label j by a “backward label” (i^-, Δ_j).

[7] LESTER RANDOLPH FORD Jr. (1927–) and DELBERT RAY FULKERSON (1924–1976), American mathematicians known for their pioneering work on flow algorithms.

If no such j exists then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 4).

4. Repeat Step 3 until t is reached.

[*This gives a flow augmenting path* P: $s \to t$.]

If it is impossible to reach t then OUTPUT f. Stop

[*f is the maximum flow.*]

Else continue (that is, go to Step 5).

5. Backtrack the path P, using the labels.

6. Using P, augment the existing flow by Δ_t. Set $f = f + \Delta_t$.

7. Remove all labels from vertices $2, \cdots, n$. Go to Step 3.

End FORD–FULKERSON

EXAMPLE 1 Ford–Fulkerson Algorithm

Applying the Ford–Fulkerson algorithm, determine the maximum flow for the network in Fig. 500 (which is the same as that in Example 1, Sec. 23.6, so that we can compare).

Solution. The algorithm proceeds as follows.

1. An initial flow $f = 9$ is given.

2. Label s $(= 1)$ by $\varnothing$. Mark 2, 3, 4, 5, 6 "unlabeled."

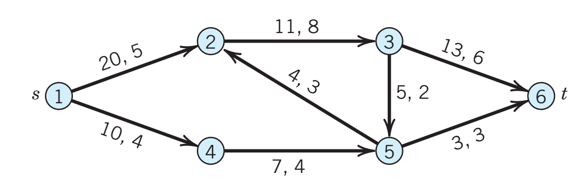


Fig. 500. Network in Example 1 with capacities (first numbers) and given flow

3. Scan 1.

Compute $\Delta_{12} = 20 - 5 = 15 = \Delta_2$. Label 2 by $(1^+, 15)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

Compute $\Delta_{23} = 11 - 8 = 3$, $\Delta_3 = \min(\Delta_2, 3) = 3$. Label 3 by $(2^+, 3)$.

Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

Scan 3.

Compute $\Delta_{36} = 13 - 6 = 7$, $\Delta_6 = \Delta_t = \min(\Delta_3, 7) = 3$. Label 6 by $(3^+, 3)$.

5. P: $1 - 2 - 3 - 6\ (= t)$ is a flow augmenting path.

6. $\Delta_t = 3$. Augmentation gives $f_{12} = 8$, $f_{23} = 11$, $f_{36} = 9$, other f_{ij} unchanged. Augmented flow $f = 9 + 3 = 12$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

3. Scan 1.

Compute $\Delta_{12} = 20 - 8 = 12 = \Delta_2$. Label 2 by $(1^+, 12)$.

Compute $\Delta_{14} = 10 - 4 = 6 = \Delta_4$. Label 4 by $(1^+, 6)$.

4. Scan 2.

Compute $\Delta_5 = \min(\Delta_2, 3) = 3$. Label 5 by $(2^-, 3)$.

Scan 4. [*No vertex left for labeling.*]

Scan 5.

Compute $\Delta_3 = \min(\Delta_5, 2) = 2$. Label 3 by $(5^-, 2)$.

Scan 3.

Compute $\Delta_{36} = 13 - 9 = 4$, $\Delta_6 = \min(\Delta_3, 4) = 2$. Label 6 by $(3^+, 2)$.

5. P: $1 - 2 - 5 - 3 - 6\ (= t)$ is a flow augmenting path.

6. $\Delta_t = 2$. Augmentation gives $f_{12} = 10, f_{32} = 1, f_{35} = 0, f_{36} = 11$, other f_{ij} unchanged. Augmented flow $f = 12 + 2 = 14$.

7. Remove labels on vertices $2, \cdots, 6$. Go to Step 3.

One can now scan 1 and then scan 2, as before, but in scanning 4 and then 5 one finds that no vertex is left for labeling. Thus one can no longer reach t. Hence the flow obtained (Fig. 501) is maximum, in agreement with our result in the last section. ■

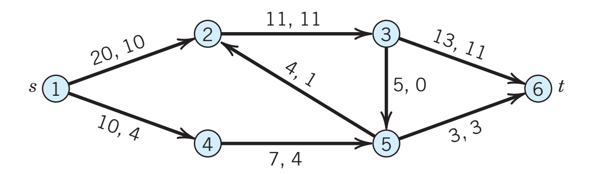


Fig. 501. Maximum flow in Example 1

PROBLEM SET 23.7

1. Do the computations indicated near the end of Example 1 in detail.
2. Solve Example 1 by Ford–Fulkerson with initial flow 0. Is it more work than in Example 1?
3. Which are the "bottleneck" edges by which the flow in Example 1 is actually limited? Hence which capacities could be decreased without decreasing the maximum flow?
4. What is the (simple) reason that Kirchhoff's law is preserved in augmenting a flow by the use of a flow augmenting path?
5. How does Ford–Fulkerson prevent the formation of cycles?

6–9 MAXIMUM FLOW

Find the maximum flow by Ford-Fulkerson:

6. In Prob. 12, Sec. 23.6
7. In Prob. 15, Sec. 23.6
8. In Prob. 14, Sec. 23.6
9.

10. **Integer flow theorem.** Prove that, if the capacities in a network G are integers, then a maximum flow exists and is an integer.
11. **CAS PROBLEM. Ford–Fulkerson.** Write a program and apply it to Probs. 6–9.
12. How can you see that Ford–Fulkerson follows a BFS technique?
13. Are the consecutive flow augmenting paths produced by Ford–Fulkerson unique?
14. If the Ford–Fulkerson algorithm stops without reaching t, show that the edges with one end labeled and the other end unlabeled form a cut set (S, T) whose capacity equals the maximum flow.
15. Find a minimum cut set in Fig. 500 and its capacity.
16. Show that in a network G with all $c_{ij} = 1$, the maximum flow equals the number of edge-disjoint paths $s \rightarrow t$.
17. In Prob. 15, the cut set contains precisely all forward edges used to capacity by the maximum flow (Fig. 501). Is this just by chance?
18. Show that in a network G with capacities all equal to 1, the capacity of a minimum cut set (S, T) equals the minimum number q of edges whose deletion destroys all directed paths $s \rightarrow t$. (A **directed path** $v \rightarrow w$ is a path in which each edge has the direction in which it is traversed in going from v to w.)

19. Several sources and sinks. If a network has several sources $s_1, \cdots, s_k$, show that it can be reduced to the case of a single-source network by introducing a new vertex s and connecting s to $s_1, \cdots, s_k$ by k edges of capacity ∞. Similarly if there are several sinks. Illustrate this idea by a network with two sources and two sinks.

20. Find the maximum flow in the network in Fig. 502 with two sources (factories) and two sinks (consumers).

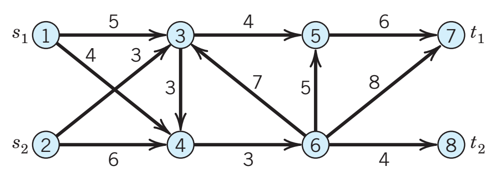


Fig. 502. Problem 20

23.8 Bipartite Graphs. Assignment Problems

From digraphs we return to graphs and discuss another important class of combinatorial optimization problems that arises in **assignment problems** of workers to jobs, jobs to machines, goods to storage, ships to piers, classes to classrooms, exams to time periods, and so on. To explain the problem, we need the following concepts.

A **bipartite graph** $G = (V, E)$ is a graph in which the vertex set V is partitioned into two sets S and T (without common elements, by the definition of a partition) such that every edge of G has one end in S and the other in T. Hence there are no edges in G that have both ends in S or both ends in T. Such a graph $G = (V, E)$ is also written $G = (S, T; E)$.

Figure 503 shows an illustration. V consists of seven elements, three workers a, b, c, making up the set S, and four jobs 1, 2, 3, 4, making up the set T. The edges indicate that worker a can do the jobs 1 and 2, worker b the jobs 1, 2, 3, and worker c the job 4. The problem is to assign one job to each worker so that every worker gets one job to do. This suggests the next concept, as follows.

DEFINITION

Maximum Cardinality Matching

A **matching** in $G = (S, T; E)$ is a set M of edges of G such that no two of them have a vertex in common. If M consists of the greatest possible number of edges, we call it a **maximum cardinality matching** in G.

For instance, a matching in Fig. 503 is $M_1 = \{(a, 2), (b, 1)\}$. Another is $M_2 = \{(a, 1), (b, 3), (c, 4)\}$; obviously, this is of maximum cardinality.

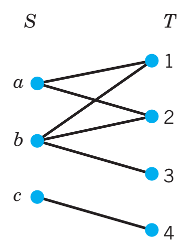


Fig. 503. Bipartite graph in the assignment of a set $S = \{a, b, c\}$ of workers to a set $T = \{1, 2, 3, 4\}$ of jobs

A vertex v is **exposed** (or *not covered*) by a matching M if v is not an endpoint of an edge of M. This concept, which always refers to some matching, will be of interest when we begin to augment given matchings (below). If a matching leaves no vertex exposed,

we call it a **complete matching**. Obviously, a complete matching can exist only if S and T consist of the same number of vertices.

We now want to show how one can stepwise increase the cardinality of a matching M until it becomes maximum. Central in this task is the concept of an augmenting path.

An **alternating path** is a path that consists alternately of edges in M and not in M (Fig. 504A). An **augmenting path** is an alternating path both of whose endpoints (a and b in Fig. 504B) are exposed. By dropping from the matching M the edges that are on an augmenting path P (two edges in Fig. 504B) and adding to M the other edges of P (three in the figure), we get a new matching, with one more edge than M. This is how we use an augmenting path in ***augmenting a given matching*** by one edge. We assert that this will always lead, after a number of steps, to a maximum cardinality matching. Indeed, the basic role of augmenting paths is expressed in the following theorem.

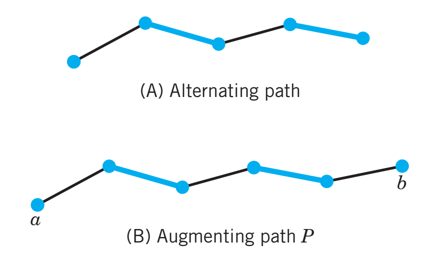


Fig. 504. Alternating and augmenting paths. Heavy edges are those belonging to a matching M

THEOREM 1

Augmenting Path Theorem for Bipartite Matching

A matching M in a bipartite graph $G = (S, T; E)$ is of maximum cardinality if and only if there does not exist an augmenting path P with respect to M.

PROOF **(a)** We show that if such a path P exists, then M is not of maximum cardinality. Let P have q edges belonging to M. Then P has $q + 1$ edges not belonging to M. (In Fig. 504B we have $q = 2$.) The endpoints a and b of P are exposed, and all the other vertices on P are endpoints of edges in M, by the definition of an alternating path. Hence if an edge of M is not an edge of P, it cannot have an endpoint on P since then M would not be a matching. Consequently, the edges of M not on P, together with the $q + 1$ edges of P not belonging to M form a matching of cardinality one more than the cardinality of M because we omitted q edges from M and added $q + 1$ instead. Hence M cannot be of maximum cardinality.

(b) We now show that if there is no augmenting path for M, then M is of maximum cardinality. Let M^* be a maximum cardinality matching and consider the graph H consisting of all edges that belong either to M or to M^*, but not to both. Then it is possible that two edges of H have a vertex in common, but three edges cannot have a vertex in common since then two of the three would have to belong to M (or to M^*), violating that M and M^* are matchings. So every v in V can be in common with two edges of H or with one or none. Hence we can characterize each "component" (= maximal *connected* subset) of H as follows.

(A) A component of H can be a closed path with an *even* number of edges (in the case of an *odd* number, two edges from M or two from M^* would meet, violating the matching property). See (A) in Fig. 505.

(B) A component of H can be an open path P with the same number of edges from M and edges from M^*, for the following reason. P must be alternating, that is, an edge of M is followed by an edge of M^*, etc. (since M and M^* are matchings). Now if P had an edge more from M^*, then P would be augmenting for M [see (B2) in Fig. 505], contradicting our assumption that there is no augmenting path for M. If P had an edge more from M, it would be augmenting for M^* [see (B3) in Fig. 505], violating the maximum cardinality of M^*, by part (a) of this proof. Hence in each component of H, the two matchings have the same number of edges. Adding to this the number of edges that belong to both M and M^* (which we left aside when we made up H), we conclude that M and M^* must have the same number of edges. Since M^* is of maximum cardinality, this shows that the same holds for M, as we wanted to prove. ■

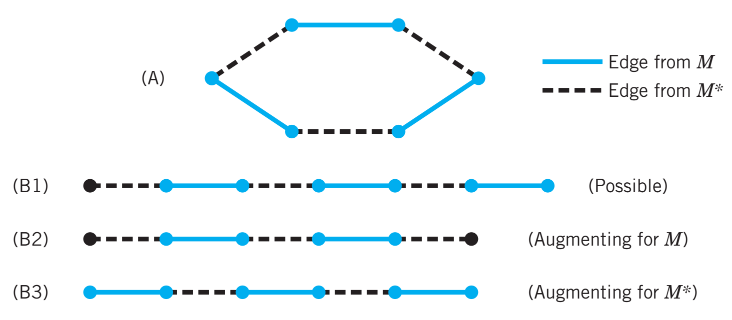


Fig. 505. Proof of the augmenting path theorem for bipartite matching

This theorem suggests the algorithm in Table 23.9 for obtaining augmenting paths, in which vertices are labeled for the purpose of backtracking paths. Such a label is *in addition* to the number of the vertex, which is also retained. Clearly, to get an augmenting path, one must start from an *exposed* vertex, and then trace an alternating path until one arrives at another *exposed* vertex. After Step 3 all vertices in S are labeled. In Step 4, the set T contains at least one exposed vertex, since otherwise we would have stopped at Step 1.

Table 23.9 Bipartite Maximum Cardinality Matching

ALGORITHM MATCHING [$G = (S, T; E), M, n$]

This algorithm determines a maximum cardinality matching M in a bipartite graph G by augmenting a given matching in G.

INPUT: Bipartite graph $G = (S, T; E)$ with vertices $1, \cdots, n$, matching M in G (for instance, $M = \varnothing$)

OUTPUT: Maximum cardinality matching M in G

1. If there is no exposed vertex in S then

OUTPUT M. Stop

[*M is of maximum cardinality in G.*]

Else label all *exposed* vertices *in* S with $\varnothing$.

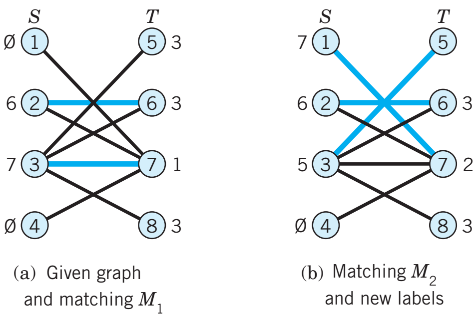
2. For each i in S and edge (i, j) *not* in M, label j with i, unless already labeled.

3. For each *nonexposed* j in T, label i with j, where i is the other end of the unique edge (i, j) in M.
4. Backtrack the alternating path P ending on an exposed vertex in T by using the labels on the vertices.
5. If no P in Step 4 is augmenting then
   OUTPUT M. Stop
   [*M is of maximum cardinality in G.*]
   Else augment M by using an augmenting path P.
   Remove all labels.
   Go to Step 1.

End MATCHING

EXAMPLE 1 Maximum Cardinality Matching

Is the matching M_1 in Fig. 506a of maximum cardinality? If not, augment it until maximum cardinality is reached.



(a) Given graph and matching M_1

(b) Matching M_2 and new labels

Fig. 506. Example 1

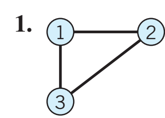
Solution. We apply the algorithm.

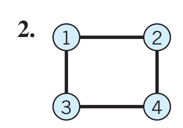
1. Label 1 and 4 with ∅.
2. Label 7 with 1. Label 5, 6, 8 with 3.
3. Label 2 with 6, and 3 with 7.
   [*All vertices are now labeled as shown in Fig.* 506a.]
4. P_1: 1 — 7 − 3 − 5. [*By backtracking*, P_1 *is augmenting*.]
   P_2: 1 — 7 − 3 − 8. [P_2 *is augmenting*.]
5. Augment M_1 by using P_1, dropping (3, 7) from M_1 and including (1, 7) and (3, 5). Remove all labels. Go to Step 1.
   Figure 506b *shows the resulting matching* $M_2 = \{(1, 7), (2, 6), (3, 5)\}$.
1. Label 4 with ∅.
2. Label 7 with 2. Label 6 and 8 with 3.
3. Label 1 with 7, and 2 with 6, and 3 with 5.
4. P_3: 5 — 3 − 8. [P_3 *is alternating but not augmenting*.]
5. Stop. M_2 *is of maximum cardinality* (*namely*, 3). ■

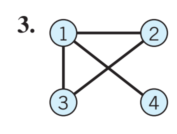
PROBLEM SET 23.8

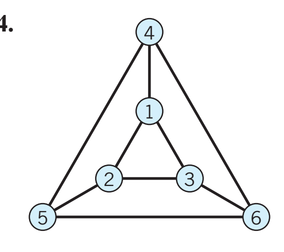
1–7 BIPARTITE OR NOT?

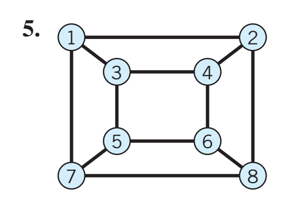
If you answer is yes, find S and T:

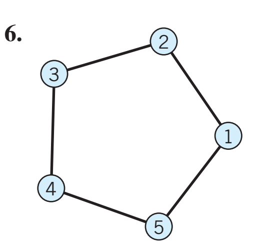
1. 

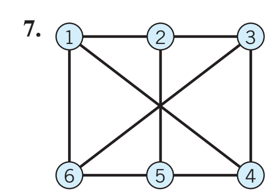
2. 

3. 

4. 

5. 

6. 

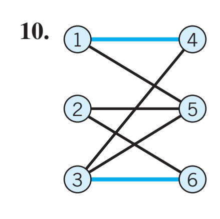
7. 

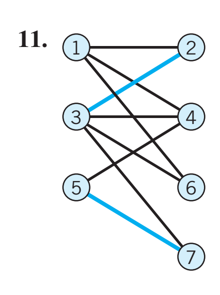
8. Can you obtain the answer to Prob. 3 from that to Prob. 1?

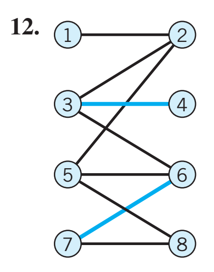
9. Can you obtain a bipartite subgraph in Prob. 4 by omitting two edges? Any two edges? Any two edges without a common vertex?

10–12 MATCHING. AUGMENTING PATHS

Find an augmenting path:

10. 

11. 

12. 

13–15 MAXIMUM CARDINALITY MATCHING

Using augmenting paths, find a maximum cardinality matching:

13. In Prob. 11

14. In Prob. 10

15. In Prob. 12

16. Complete bipartite graphs. A bipartite graph $G = (S, T; E)$ is called *complete* if every vertex in S is joined to every vertex in T by an edge, and is denoted by K_{n_1,n_2}, where n_1 and n_2 are the numbers of vertices in S and T, respectively. How many edges does this graph have?

17. Planar graph. A *planar graph* is a graph that can be drawn on a sheet of paper so that no two edges cross. Show that the complete graph K_4 with four vertices is planar. The complete graph K_5 with five vertices is not planar. Make this plausible by attempting to draw K_5 so that no edges cross. Interpret the result in terms of a net of roads between five cities.

18. Bipartite graph $K_{3,3}$ not planar. Three factories 1, 2, 3 are each supplied underground by water, gas, and electricity, from points A, B, C, respectively. Show that this can be represented by $K_{3,3}$ (the complete bipartite graph $G = (S, T; E)$ with S and T consisting of three vertices each) and that eight of the nine supply lines (edges) can be laid out without crossing. Make it plausible that $K_{3,3}$ is not planar by attempting to draw the ninth line without crossing the others.

19–25 VERTEX COLORING

19. Vertex coloring and exam scheduling. What is the smallest number of exam periods for six subjects a, b, c, d, e, f if some of the students simultaneously take a, b, f, some c, d, e, some a, c, e, and some c, e? Solve this as follows. Sketch a graph with six vertices $a, \cdots, f$ and join vertices if they represent subjects simultaneously taken by some students. Color the vertices so that adjacent vertices receive different colors. (Use numbers $1, 2, \cdots$ instead of actual colors if you want.) What is the minimum number of colors you need? For any graph G, this minimum number is called the

(vertex) **chromatic number** $\chi_v(G)$. Why is this the answer to the problem? Write down a possible schedule.

20. Scheduling and matching. Three teachers x_1, x_2, x_3 teach four classes y_1, y_2, y_3, y_4 for these numbers of periods:

	y_1	y_2	y_3	y_4
x_1	1	0	1	1
x_2	1	1	1	1
x_3	0	1	1	1

Show that this arrangement can be represented by a bipartite graph G and that a teaching schedule for one period corresponds to a matching in G. Set up a teaching schedule with the smallest possible number of periods.

21. How many colors do you need for vertex coloring any tree?

22. Harbor management. How many piers does a harbor master need for accommodating six cruise ships $S_1, \cdots, S_6$ with expected dates of arrival A and departure D in July, $(A, D) = (10, 13)$, $(13, 15)$, $(14, 17)$, $(12, 15)$, $(16, 18)$, $(14, 17)$, respectively, if each pier can accommodate only one ship, arrival being at 6 am and departures at 11 pm? *Hint.* Join S_i and S_j by an edge if their intervals overlap. Then color vertices.

23. What would be the answer to Prob. 22 if only the five ships $S_1, \cdots, S_5$ had to be accommodated?

24. Four- (vertex) color theorem. The famous *four-color theorem* states that one can color the vertices of any *planar* graph (so that adjacent vertices get different colors) with at most four colors. It had been conjectured for a long time and was eventually proved in 1976 by Appel and Haken [*Illinois J. Math* **21** (1977), 429–567]. Can you color the complete graph K_5 with four colors? Does the result contradict the four-color theorem? (For more details, see Ref. [F1] in App. 1.)

25. Find a graph, as simple as possible, that cannot be vertex colored with three colors. Why is this of interest in connection with Prob. 24?

26. Edge coloring. The *edge chromatic number* $\chi_e(G)$ of a graph G is the minimum number of colors needed for coloring the edges of G so that incident edges get different colors. Clearly, $\chi_e(G) \geqq \max d(u)$, where $d(u)$ is the degree of vertex u. If $G = (S, T; E)$ is bipartite, the equality sign holds. Prove this for $K_{n,n}$ the complete (cf. Sec. 23.1) bipartite graph $G = (S, T, E)$ with S and T consisting of n vertices each.

CHAPTER 23 REVIEW QUESTIONS AND PROBLEMS

1. What is a graph, a digraph, a cycle, a tree?

2. State some typical problems that can be modeled and solved by graphs or digraphs.

3. State from memory how graphs can be handled on computers.

4. What is a shortest path problem? Give applications.

5. What situations can be handled in terms of the traveling salesman problem?

6. Give typical applications involving spanning trees.

7. What are the basic ideas and concepts in handling flows?

8. What is combinatorial optimization? Which sections of this chapter involved it? Explain details.

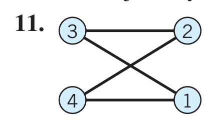
9. Define bipartite graphs and describe some typical applications of them.

10. What is BFS? DFS? In what connection did these concepts occur?

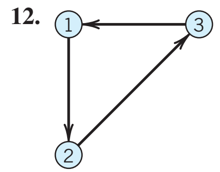
11–16 MATRICES FOR GRAPHS AND DIGRAPHS

Find the adjacency matrix of:

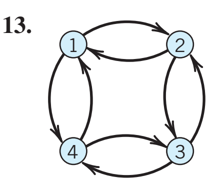
11.



12.



13.



14–16 Sketch the graph whose adjacency matrix is:

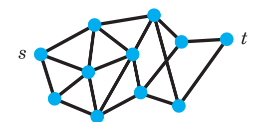
14. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

15. $\begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

16. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{bmatrix}$

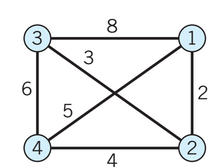
17. Vertex incidence list. Make it for the graph in Prob. 15.

18. Find a shortest path and its length by Moore's BFS algorithm, assuming that all the edges have length 1.



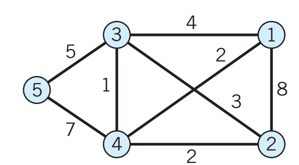
Problem 18

19. Find shortest paths by Dijkstra's algorithm.



Problem 19

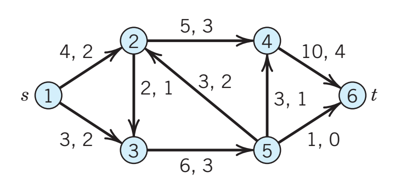
20. Find a shortest spanning tree.



Problem 20

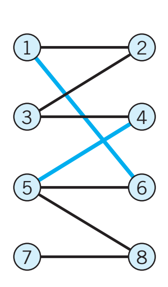
21. Company A has offices in Chicago, Los Angeles, and New York; Company B in Boston and New York; Company C in Chicago, Dallas, and Los Angeles. Represent this by a bipartite graph.

22. Find flow augmenting paths and the maximum flow.



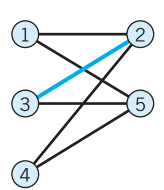
Problem 22

23. Using augmenting paths, find a maximum cardinality matching.



Problem 25

24. Find an augmenting path,



Problem 24

SUMMARY OF CHAPTER 23
Graphs. Combinatorial Optimization

Combinatorial optimization concerns optimization problems of a discrete or combinatorial structure. It uses graphs and digraphs (Sec. 23.1) as basic tools.

A **graph** $G = (V, E)$ consists of a set V of **vertices** $v_1, v_2, \cdots, v_n$ (often simply denoted by $1, 2, \cdots, n$) and a set E of **edges** $e_1, e_2, \cdots, e_m$, each of which connects two vertices. We also write (i, j) for an edge with vertices i and j as endpoints. A **digraph** (= directed graph) is a graph in which each edge has a direction (indicated by an arrow). For handling graphs and digraphs in computers, one can use *matrices* or *lists* (Sec. 23.1).

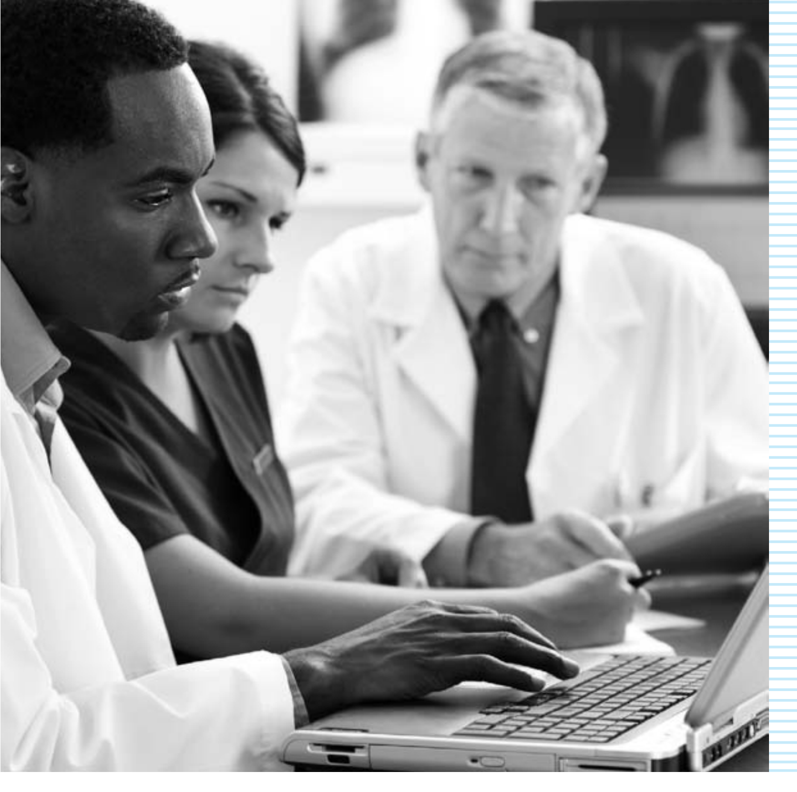
This chapter is devoted to important classes of optimization problems for graphs and digraphs that all arise from practical applications, and corresponding algorithms, as follows.

In a **shortest path problem** (Sec. 23.2) we determine a path of minimum length (consisting of edges) from a vertex s to a vertex t in a graph whose edges (i, j) have a "length" $l_{ij} > 0$, which may be an actual length or a travel time or cost or an electrical resistance [if (i, j) is a wire in a net], and so on. ***Dijkstra's algorithm*** (Sec. 23.3) or, when all $l_{ij} = 1$, ***Moore's algorithm*** (Sec. 23.2) are suitable for these problems.

A **tree** is a graph that is connected and has no **cycles** (no closed paths). Trees are very important in practice. A *spanning tree* in a graph G is a tree containing *all* the vertices of G. If the edges of G have lengths, we can determine a **shortest spanning tree**, for which the sum of the lengths of all its edges is minimum, by ***Kruskal's algorithm*** or ***Prim's algorithm*** (Secs. 23.4, 23.5).

A **network** (Sec. 23.6) is a digraph in which each edge (i, j) has a *capacity* $c_{ij} > 0$ [= maximum possible flow along (i, j)] and at one vertex, the *source* s, a flow is produced that flows along the edges to a vertex t, the *sink* or *target,* where the flow disappears. The problem is to maximize the flow, for instance, by applying the **Ford–Fulkerson algorithm** (Sec. 23.7), which uses ***flow augmenting paths*** (Sec. 23.6). Another related concept is that of a ***cut set***, as defined in Sec. 23.6.

A **bipartite graph** $G = (V, E)$ (Sec. 23.8) is a graph whose vertex set V consists of two parts S and T such that every edge of G has one end in S and the other in T, so that there are no edges connecting vertices in S or vertices in T. A **matching** in G is a set of edges, no two of which have an endpoint in common. The problem then is to find a **maximum cardinality matching** in G, that is, a matching M that has a maximum number of edges. For an algorithm, see Sec. 23.8.



PART G

Probability, Statistics

Probability theory (Chap. 24) provides models of probability distributions (theoretical models of the observable reality involving chance effects) to be tested by statistical methods, and it will also supply the mathematical foundation of these methods in Chap. 25.

Modern **mathematical statistics** (Chap. 25) has various engineering applications, for instance, in testing materials, control of production processes, quality control of production outputs, performance tests of systems, robotics, and automatization in general, production planning, marketing analysis, and so on.

To this we could add a long list of fields of applications, for instance, in agriculture, biology, computer science, demography, economics, geography, management of natural resources, medicine, meteorology, politics, psychology, sociology, traffic control, urban planning, etc. Although these applications are very heterogeneous, we shall see that most statistical methods are universal in the sense that each of them can be applied in various fields.

Additional Software for Probability and Statistics

See also the list of software at the beginning of Part E on Numerical Analysis.

Data Desk. Data Description, Inc., Ithaca, NY. Phone 1-800-573-5121 or (607) 257-1000, website at www.datadesk.com.

MINITAB. Minitab, Inc., State College, PA. Phone 1-800-448-3555 or (814) 238-3280, website at www.minitab.com.

SAS. SAS Institute, Inc., Cary, NC. Phone 1-800-727-0025 or (919) 677-8000, website at www.sas.com.

R. website at www.r-project.org. Free software, part of the GNU/Free Software Foundation project.

SPSS. SPSS, Inc., Chicago, IL. (part of IBM) Phone 1-800-543-2185 or (312) 651-3000, website at www.spss.com.

STATISTICA. StatSoft, Inc., Tulsa, OK. Phone (918) 749-1119, website at www.statsoft.com.

TIBCO Spotfire S+. TIBCO Software Inc., Palo Alto, CA; Office for this software: Somerville, MA. Phone 1-866-240-0491 (toll-free), (617) 702-1602, website at spotfire.tibco.com/products/s-plus/statistical-analysis-software.aspx

CHAPTER 24

Data Analysis. Probability Theory

We first show how to handle data numerically or in terms of graphs, and how to extract information (average size, spread of data, etc.) from them. If these data are influenced by "chance," by factors whose effect we cannot predict exactly (e.g., weather data, stock prices, life spans of tires, etc.), we have to rely on **probability theory**. This theory originated in games of chance, such as flipping coins, rolling dice, or playing cards. Nowadays it gives mathematical models of chance processes called *random experiments* or, briefly, **experiments**. In such an experiment we observe a **random variable** X, that is, a function whose values in a **trial** (a performance of an experiment) occur "by chance" (Sec. 24.3) according to a **probability distribution** that gives the individual probabilities with which possible values of X may occur in the long run. (Example: Each of the six faces of a die should occur with the same probability, $1/6$.) Or we may simultaneously observe more than one random variable, for instance, height *and* weight of persons or hardness *and* tensile strength of steel. This is discussed in Sec. 24.9, which will also give the basis for the mathematical justification of the statistical methods in Chapter 25.

Prerequisite: Calculus.
References and Answers to Problems: App. 1 Part G, App. 2.

24.1 Data Representation. Average. Spread

Data can be represented numerically or graphically in various ways. For instance, your daily newspaper may contain tables of stock prices and money exchange rates, curves or bar charts illustrating economical or political developments, or pie charts showing how your tax dollar is spent. And there are numerous other representations of data for special purposes.

In this section we discuss the use of standard representations of data in statistics. (For these, software packages, such as DATA DESK, R, and MINITAB, are available, and Maple or Mathematica may also be helpful; see pp. 789 and 1009) We explain corresponding concepts and methods in terms of typical examples.

EXAMPLE 1 Recording and Sorting

Sample values (observations, measurements) should be **recorded** in the order in which they occur. **Sorting**, that is, ordering the sample values by size, is done as a first step of investigating properties of the sample and graphing it. Sorting is a standard process on the computer; see Ref. [E35], listed in App. 1.

Super alloys is a collective name for alloys used in jet engines and rocket motors, requiring high temperature (typically 1800°F), high strength, and excellent resistance to oxidation. Thirty specimens of Hastelloy C (nickel-based steel, investment cast) had the tensile strength (in 1000 lb/sq in.), recorded in the order obtained and rounded to integer values,

(1)	89	77	88	91	88	93	99	79	87	84	86	82	88	89	78
	90	91	81	90	83	83	92	87	89	86	89	81	87	84	89

Sorting gives

(2)	77	78	79	81	81	82	83	83	84	84	86	86	87	87	87
	88	88	88	89	89	89	89	89	90	90	91	91	92	93	99

■

Graphic Representation of Data

We shall now discuss standard graphic representations used in statistics for obtaining information on properties of data.

EXAMPLE 2 **Stem-and-Leaf Plot (Fig. 507)**

This is one of the simplest but most useful representations of data. For (1) it is shown in Fig. 507. The numbers in (1) range from 78 to 99; see (2). We divide these numbers into 5 groups, 75–79, 80–84, 85–89, 90–94, 95–99. The integers in the tens position of the groups are 7, 8, 8, 9, 9. These form the *stem* in Fig. 507. The first *leaf* is 789, representing 77, 78, 79. The second leaf is 1123344, representing 81, 81, 82, 83, 83, 84, 84. And so on.

The number of times a value occurs is called its **absolute frequency**. Thus 78 has absolute frequency 1, the value 89 has absolute frequency 5, etc. The column to the extreme left in Fig. 507 shows the **cumulative absolute frequencies**, that is, the sum of the absolute frequencies of the values up to the line of the leaf. Thus, the number 10 in the second line on the left shows that (1) has 10 values up to and including 84. The number 23 in the next line shows that there are 23 values not exceeding 89, etc. Dividing the cumulative absolute frequencies by n (= 30 in Fig. 507) gives the **cumulative relative frequencies** 0.1, 0.33, 0.76, 0.93, 1.00. ■

EXAMPLE 3 **Histogram (Fig. 508)**

For large sets of data, histograms are better in displaying the distribution of data than stem-and-leaf plots. The principle is explained in Fig. 508. (An application to a larger data set is shown in Sec. 25.7). The bases of the rectangles in Fig. 508 are the x-intervals (known as **class intervals**) 74.5–79.5, 79.5–84.5, 84.5–89.5, 89.5–94.5, 94.5–99.5, whose midpoints (known as **class marks**) are $x = 77, 82, 87, 92, 97$, respectively. The height of a rectangle with class mark x is the **relative class frequency** $f_{\text{rel}}(x)$, defined as the number of data values in that class interval, divided by n (= 30 in our case). Hence the areas of the rectangles are proportional to these relative frequencies, 0.10, 0.23, 0.43, 0.17, 0.07, so that histograms give a good impression of the distribution of data. ■

Leaf unit = 1.0

3	7	789
10	8	1123344
23	8	6677788899999
29	9	001123
30	9	9

Fig. 507. Stem-and-leaf plot of the data in Example 1

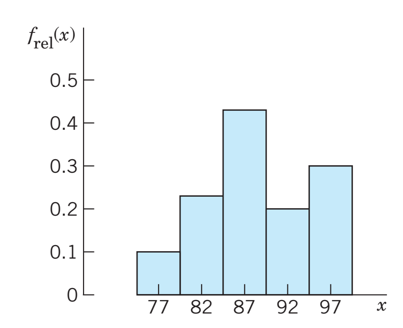


Fig. 508. Histogram of the data in Example 1 (grouped as in Fig. 507)

EXAMPLE 4 Boxplot. Median. Interquartile Range. Outlier

A **boxplot** of a set of data illustrates the average size and the spread of the values, in many cases the two most important quantities characterizing the set, as follows.

The average size is measured by the **median**, or *middle quartile,* q_M. If the number n of values of the set is *odd*, then q_M is the middlemost of the values when ordered as in (2). If n is *even,* then q_M is the average of the two middlemost values of the ordered set. In (2) we have $n = 30$ and thus $q_M = \frac{1}{2}(x_{15} + x_{16}) = \frac{1}{2}(87 + 88) = 87.5$. (In general, q_M will be a fraction if n is even.)

The spread of values can be measured by the **range** $R = x_{\max} - x_{\min}$, the largest value minus the smallest one.

Better information on the spread gives the **interquartile range** $\text{IQR} = q_U - q_L$. Here q_U is the middlemost value (or the average of the two middlemost values) in the data *above* the median; and q_L is the middlemost value (or the average of the two middlemost values) in the data *below* the median. Hence in (2) we have $q_U = x_{23} = 89$, $q_L = x_8 = 83$, and $\text{IQR} = 89 - 83 = 6$.

The box in Fig. 509 extends vertically from q_L to q_U; it has height $\text{IQR} = 6$. The vertical lines below and above the box extend from $x_{\min} = 77$ to $x_{\max} = 99$, so that they show $R = 22$.

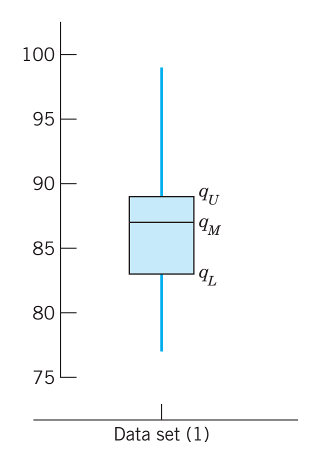


Fig. 509. Boxplot of the data set (1)

The line above the box is suspiciously long. This suggests the concept of an **outlier**, a value that is more than 1.5 times the IQR away from either end of the box; here 1.5 is purely conventional. An outlier indicates that something might have gone wrong in the data collection. In (2) we have $89 + 1.5\ \text{IQR} = 98$, and we regard 99 as an outlier. ■

Mean. Standard Deviation. Variance. Empirical Rule

Medians and quartiles are easily obtained by ordering and counting, practically without calculation. But they do not give full information on data: you can change data values to some extent without changing the median. Similarly for the quartiles.

The average size of the data values can be measured in a more refined way by the **mean**

(3)
$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n).$$

This is the arithmetic mean of the data values, obtained by taking their sum and dividing by the data *size* n. Thus in (1),

$$\bar{x} = \tfrac{1}{30}(89 + 77 + \cdots + 89) = \tfrac{260}{3} \approx 86.7.$$

Every data value contributes, and changing one of them will change the mean.

Similarly, the spread (variability) of the data values can be measured in a more refined way by the **standard deviation** s or by its square, the **variance**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2]. \tag{4}$$

Thus, to obtain the variance of the data, take the difference $x_j - \bar{x}$ of each data value from the mean, square it, take the sum of these n squares, and divide it by $n - 1$ (not n, as we motivate in Sec. 25.2). To get the standard deviation s, take the square root of s^2.

For example, using $\bar{x} = 260/3$, we get for the data (1) the variance

$$s^2 = \tfrac{1}{29}[(89 - \tfrac{260}{3})^2 + (77 - \tfrac{260}{3})^2 + \cdots + (89 - \tfrac{260}{3})^2] = \tfrac{2006}{87} \approx 23.06$$

Hence the standard deviation is $s = \sqrt{2006/87} \approx 4.802$. Note that the standard deviation has the same dimension as the data values (kg/mm^2, see at the beginning), which is an advantage. On the other hand, the variance is preferable to the standard deviation in developing statistical methods, as we shall see in Chap. 25.

CAUTION! Your CAS (Maple, for instance) may use $1/n$ instead of $1/(n - 1)$ in (4), but the latter is better when n is small (see Sec. 25.2).

Mean and standard deviation, introduced to give center and spread, actually give much more information according to this rule.

Empirical Rule. For any mound-shaped, nearly symmetric distribution of data the intervals

$$\bar{x} \pm s, \quad \bar{x} \pm 2s, \quad \bar{x} \pm 3s \qquad \text{contain about} \qquad 68\%, \quad 95\%, \quad 99.7\%,$$

respectively, of the data points.

EXAMPLE 5 Empirical Rule and Outliers. *z*-Score

For (1), with $\bar{x} = 86.7$ and $s = 4.8$, the three intervals in the Rule are $81.9 \leqq x \leqq 91.5$, $77.1 \leqq x \leqq 96.3$, $72.3 \leqq x \leqq 101.1$ and contain 73% (22 values remain, 5 are too small, and 5 too large), 93% (28 values, 1 too small, and 1 too large), and 100%, respectively.

If we reduce the sample by omitting the outlier 99, mean and standard deviation reduce to $\bar{x}_{\text{red}} = 86.2$, $s_{\text{red}} = 4.3$, approximately, and the percentage values become 67% (5 and 5 values outside), 93% (1 and 1 outside), and 100%.

Finally, the relative position of a value x in a set of mean $\bar{x}$ and standard deviation s can be measured by the **z-score**

$$z(s) = \frac{x - \bar{x}}{s}.$$

This is the distance of x from the mean $\bar{x}$ measured in multiples of s. For instance, $z(83) = (83 - 86.7)/4.8 = -0.77$. This is negative because 83 lies below the mean. By the Empirical Rule, the extreme z-values are about -3 and 3. ■

PROBLEM SET 24.1

1–10 DATA REPRESENTATIONS

Represent the data by a stem-and-leaf plot, a histogram, and a boxplot:

1. Length of nails [mm]

19 21 19 20 19 20 21 20

2. Phone calls per minute in an office between 9:00 A.M. and 9:10 A.M.

6 6 4 2 1 7 0 4 6 7

3. Systolic blood pressure of 15 female patients of ages 20–22

156 158 154 133 141 130 144 137
151 146 156 138 138 149 139

4. Iron content [%] of 15 specimens of hermatite (Fe_2O_3)

72.8 70.4 71.2 69.2 70.3 68.9 71.1 69.8
71.5 69.7 70.5 71.3 69.1 70.9 70.6

5. Weight of filled bags [g] in an automatic filling

203 199 198 201 200 201 201

6. Gasoline consumption [miles per gallon, rounded] of six cars of the same model under similar conditions

15.0 15.5 14.5 15.0 15.5 15.0

7. Release time [sec] of a relay

1.3 1.2 1.4 1.5 1.3 1.3 1.4 1.1 1.5 1.4
1.6 1.3 1.5 1.1 1.4 1.2 1.3 1.5 1.4 1.4

8. Foundrax test of Brinell hardness (2.5 mm steel ball, 62.5 kg load, 30 sec) of 20 copper plates (values in kg/mm^2)

86 86 87 89 76 85 82 86 87 85
90 88 89 90 88 80 84 89 90 89

9. Efficiency [%] of seven Voith Francis turbines of runner diameter 2.3 m under a head range of 185 m

91.8 89.1 89.9 92.5 90.7 91.2 91.0

10. −0.51 0.12 −0.47 0.95 0.25 −0.18 −0.54

11–16 AVERAGE AND SPREAD

Find the mean and compare it with the median. Find the standard deviation and compare it with the interquartile range.

11. For the data in Prob. 1

12. For the phone call data in Prob. 2

13. For the medical data in Prob. 3

14. For the iron contents in Prob. 4

15. For the release times in Prob. 7

16. For the Brinell hardness data in Prob. 8

17. Outlier, reduced data. Calculate s for the data 4 1 3 10 2. Then reduce the data by deleting the outlier and calculate s. Comment.

18. Outlier, reduction. Do the same tasks as in Prob. 17 for the hardness data in Prob. 8.

19. Construct the simplest possible data with $\bar{x} = 100$ but $q_M = 0$. What is the point of this problem?

20. Mean. Prove that $\bar{x}$ must always lie between the smallest and the largest data values.

24.2 Experiments, Outcomes, Events

We now turn to **probability theory**. This theory has the purpose of providing mathematical models of situations affected or even governed by "chance effects," for instance, in weather forecasting, life insurance, quality of technical products (computers, batteries, steel sheets, etc.), traffic problems, and, of course, games of chance with cards or dice. And the accuracy of these models can be tested by suitable observations or experiments—this is a main purpose of **statistics** to be explained in Chap. 25.

We begin by defining some standard terms. An **experiment** is a process of measurement or observation, in a laboratory, in a factory, on the street, in nature, or wherever; so "experiment" is used in a rather general sense. Our interest is in experiments that involve **randomness**, chance effects, so that we cannot predict a result exactly. A **trial** is a single performance of an experiment. Its result is called an **outcome** or a **sample point**. n trials then give a **sample** of **size** n consisting of n sample points. The **sample space** S of an experiment is the set of all possible outcomes.

EXAMPLES 1–6 Random Experiments. Sample Spaces

(1) Inspecting a lightbulb. $S = \{\text{Defective, Nondefective}\}$.

(2) Rolling a die. $S = \{1, 2, 3, 4, 5, 6\}$.

(3) Measuring tensile strength of wire. S the numbers in some interval.

(4) Measuring copper content of brass. S: 50% to 90%, say.

(5) Counting daily traffic accidents in New York. S the integers in some interval.

(6) Asking for opinion about a new car model. $S = \{\text{Like, Dislike, Undecided}\}$. ■

The subsets of S are called **events** and the outcomes **simple events**.

EXAMPLE 7 Events

In (2), events are $A = \{1, 3, 5\}$ (*"Odd number"*), $B = \{2, 4, 6\}$ (*"Even number"*), $C = \{5, 6\}$. etc. Simple events are $\{1\}, \{2\}, \cdots, \{6\}$. ■

If, in a trial, an outcome a happens and $a \in A$ (a *is an element of* A), we say that A happens. For instance, if a die turns up a 3, the event A*: Odd number* happens. Similarly, if C in Example 7 happens (meaning 5 or 6 turns up), then, say, $D = \{4, 5, 6\}$ happens. Also note that S happens in each trial, meaning that *some* event of S always happens. All this is quite natural.

Unions, Intersections, Complements of Events

In connection with basic probability laws we shall need the following concepts and facts about events (subsets) $A, B, C, \cdots$ of a given sample space S.

The **union** $A \cup B$ of A and B consists of all points in A or B or both.

The **intersection** $A \cap B$ of A and B consists of all points that are in both A and B.

If A and B have no points in common, we write

$$A \cap B = \varnothing$$

where $\varnothing$ is the *empty set* (set with no elements) and we call A and B **mutually exclusive** (or **disjoint**) because, in a trial, the occurrence of A *excludes* that of B (and conversely)—if your die turns up an odd number, it cannot turn up an even number in the same trial. Similarly, a coin cannot turn up *Head* and *Tail* at the same time.

Complement A^c of A. This is the set of all the points of S ***not*** in A. Thus,

$$A \cap A^c = \varnothing, \qquad A \cup A^c = S.$$

In Example 7 we have $A^c = B$, hence $A \cup A^c = \{1, 2, 3, 4, 5, 6\} = S$.

Another notation for the complement of A is $\bar{A}$ (instead of A^c), but we shall not use this because in set theory $\bar{A}$ is used to denote the *closure* of A (not needed in our work).

Unions and intersections of more events are defined similarly. The **union**

$$\bigcup_{j=1}^{m} A_j = A_1 \cup A_2 \cup \cdots \cup A_m$$

of events $A_1, \cdots, A_m$ consists of all points that are in at least one A_j. Similarly for the union $A_1 \cup A_2 \cup \cdots$ of infinitely many subsets $A_1, A_2, \cdots$ of an *infinite* sample space S (that is, S consists of infinitely many points). The **intersection**

$$\bigcap_{j=1}^{m} A_j = A_1 \cap A_2 \cap \cdots \cap A_m$$

of $A_1, \cdots, A_m$ consists of the points of S that are in each of these events. Similarly for the intersection $A_1 \cap A_2 \cap \cdots$ of infinitely many subsets of S.

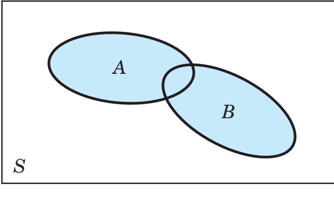
Working with events can be illustrated and facilitated by **Venn diagrams**[1] for showing unions, intersections, and complements, as in Figs. 510 and 511, which are typical examples that give the idea.

EXAMPLE 8 Unions and Intersections of 3 Events

In rolling a die, consider the events

A: *Number greater than* 3, B: *Number less than* 6, C: *Even number*.

Then $A \cap B = \{4, 5\}$, $B \cap C = \{2, 4\}$, $C \cap A = \{4, 6\}$, $A \cap B \cap C = \{4\}$. Can you sketch a Venn diagram of this? Furthermore, $A \cup B = S$, hence $A \cup B \cup C = S$ (why?). ■



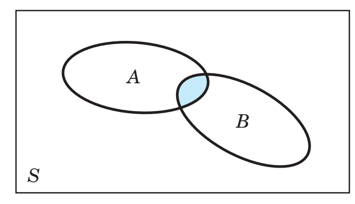


Fig. 510. Venn diagrams showing two events A and B in a sample space S and their union $A \cup B$ (colored) and intersection $A \cap B$ (colored)

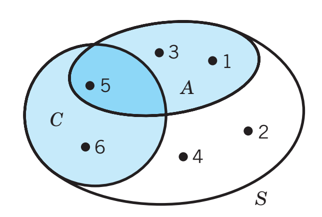


Fig. 511. Venn diagram for the experiment of rolling a die, showing S, $A = \{1, 3, 5\}$, $C = \{5, 6\}$, $A \cup C = \{1, 3, 5, 6\}$, $A \cap C = \{5\}$

PROBLEM SET 24.2

1–12 SAMPLE SPACES, EVENTS

Graph a sample space for the experiments:

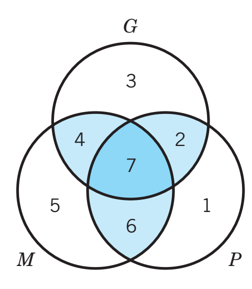
1. Drawing 3 screws from a lot of right-handed and left-handed screws
2. Tossing 2 coins
3. Rolling 2 dice
4. Rolling a die until the first *Six* appears
5. Tossing a coin until the first *Head* appears
6. Recording the lifetime of each of 3 lightbulbs

[1]JOHN VENN (1834–1923), English mathematician.

7. Recording the daily maximum temperature X and the daily maximum air pressure Y at Times Square in New York
8. Choosing a committee of 2 from a group of 5 people
9. Drawing gaskets from a lot of 10, containing one defective D, unitil D is drawn, one at a time and assuming **sampling without replacement**, that is, gaskets drawn are *not* returned to the lot. (More about this in Sec. 24.6)
10. In rolling 3 dice, are the events A: *Sum divisible by* 3 and B: *Sum divisible by* 5 mutually exclusive?
11. Answer the questions in Prob. 10 for rolling 2 dice.
12. List all 8 subsets of the sample space $S = \{a, b, c\}$.
13. In Prob. 3 circle and mark the events A: *Faces are equal*, B: *Sum of faces less than* 5, $A \cup B$, $A \cap B$, A^c, B^c.
14. In drawing 2 screws from a lot of right-handed and left-handed screws, let A, B, C, D mean at a least 1 right-handed, at least 1 left-handed, 2 right-handed, 2 left-handed, respectively. Are A and B mutually exclusive? C and D?

15–20 VENN DIAGRAMS

15. In connection with a trip to Europe by some students, consider the events P that they see Paris, G that they have a good time, and M that they run out of money, and describe in words the events $1, \cdots, 7$ in the diagram.



Problem 15

16. Show that, by the definition of complement, for any subset A of a sample space S.

$$(A^c)^c = A, \qquad S^c = \varnothing, \qquad \varnothing^c = S,$$
$$A \cup A^c = S, \qquad A \cap A^c = \varnothing.$$

17. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cup B = B$.
18. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.
19. (**De Morgan's laws**) Using Venn diagrams, graph and check *De Morgan's laws*

$$(A \cup B)^c = A^c \cap B^c$$
$$(A \cap B)^c = A^c \cup B^c.$$

20. Using Venn diagrams, graph and check the rules

$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C).$$

24.3 Probability

The "probability" of an event A in an experiment is supposed to measure how frequently A is *about* to occur if we make many trials. If we flip a coin, then heads H and tails T will appear *about* equally often—we say that H and T are "**equally likely**." Similarly, for a regularly shaped die of homogeneous material ("**fair die**") each of the six outcomes $1, \cdots, 6$ will be equally likely. These are examples of experiments in which the sample space S consists of finitely many outcomes (points) that for reasons of some symmetry can be regarded as equally likely. This suggests the following definition.

DEFINITION 1

First Definition of Probability

If the sample space S of an experiment consists of finitely many outcomes (points) that are equally likely, then the probability $P(A)$ of an event A is

(1) $$P(A) = \frac{\text{Number of points in } A}{\text{Number of points in } S}.$$

From this definition it follows immediately that, in particular,

$$P(S) = 1. \tag{2}$$

EXAMPLE 1 **Fair Die**

In rolling a fair die once, what is the probability $P(A)$ of A of obtaining a 5 or a 6? The probability of B: "*Even number*"?

Solution. The six outcomes are equally likely, so that each has probability 1/6. Thus $P(A) = 2/6 = 1/3$ because $A = \{5, 6\}$ has 2 points, and $P(B) = 3/6 = 1/2$. ■

Definition 1 takes care of many games as well as some practical applications, as we shall see, but certainly not of all experiments, simply because in many problems we do not have finitely many equally likely outcomes. To arrive at a more general definition of probability, we regard ***probability as the counterpart of relative frequency***. Recall from Sec. 24.1 that the **absolute frequency** $f(A)$ of an event A in n trials is the number of times A occurs, and the **relative frequency** of A in these trials is $f(A)/n$; thus

$$f_{\text{rel}}(A) = \frac{f(A)}{n} = \frac{\text{Number of times } A \text{ occurs}}{\text{Number of trials}}. \tag{3}$$

Now if A did not occur, then $f(A) = 0$. If A always occurred, then $f(A) = n$. These are the extreme cases. Division by n gives

$$0 \leqq f_{\text{rel}}(A) \leqq 1. \tag{4*}$$

In particular, for $A = S$ we have $f(S) = n$ because S always occurs (meaning that some event always occurs; if necessary, see Sec. 24.2, after Example 7). Division by n gives

$$f_{\text{rel}}(S) = 1. \tag{5*}$$

Finally, if A and B are mutually exclusive, they cannot occur together. Hence the absolute frequency of their union $A \cup B$ must equal the sum of the absolute frequencies of A and B. Division by n gives the same relation for the relative frequencies,

$$f_{\text{rel}}(A \cup B) = f_{\text{rel}}(A) + f_{\text{rel}}(B) \qquad (A \cap B = \varnothing). \tag{6*}$$

We are now ready to extend the definition of probability to experiments in which equally likely outcomes are not available. Of course, the extended definition should include Definition 1. Since probabilities are supposed to be the theoretical counterpart of relative frequencies, we choose the properties in (4*), (5*), (6*) as axioms. (Historically, such a choice is the result of a long process of gaining experience on what might be best and most practical.)

DEFINITION 2

General Definition of Probability

Given a sample space S, with each event A of S (subset of S) there is associated a number $P(A)$, called the **probability** of A, such that the following **axioms of probability** are satisfied.

1. For every A in S,

$$0 \leqq P(A) \leqq 1. \tag{4}$$

2. The entire sample space S has the probability

$$P(S) = 1. \tag{5}$$

3. For mutually exclusive events A and B ($A \cap B = \varnothing$; see Sec. 24.2),

$$P(A \cup B) = P(A) + P(B) \qquad (A \cap B = \varnothing). \tag{6}$$

If S is infinite (has infinitely many points), Axiom 3 has to be replaced by

3′. For mutually exclusive events $A_1, A_2, \cdots$,

$$P(A_1 \cup A_2 \cup \cdots) = P(A_1) + P(A_2) + \cdots. \tag{6′}$$

In the infinite case the subsets of S on which $P(A)$ is defined are restricted to form a so-called σ-*algebra*, as explained in Ref. [GenRef6] (not [G6]!) in App. 1. This is of no practical consequence to us.

Basic Theorems of Probability

We shall see that the axioms of probability will enable us to build up probability theory and its application to statistics. We begin with three basic theorems. The first of them is useful if we can get the probability of the complement A^c more easily than $P(A)$ itself.

THEOREM 1

Complementation Rule

For an event A and its complement A^c in a sample space S,

$$P(A^c) = 1 - P(A). \tag{7}$$

PROOF By the definition of complement (Sec. 24.2), we have $S = A \cup A^c$ and $A \cap A^c = \varnothing$. Hence by Axioms 2 and 3,

$$1 = P(S) = P(A) + P(A^c), \qquad \text{thus} \qquad P(A^c) = 1 - P(A). \qquad \blacksquare$$

EXAMPLE 2 **Coin Tossing**

Five coins are tossed simultaneously. Find the probability of the event A: *At least one head turns up.* Assume that the coins are fair.

Solution. Since each coin can turn up heads or tails, the sample space consists of $2^5 = 32$ outcomes. Since the coins are fair, we may assign the same probability (1/32) to each outcome. Then the event A^c (*No heads turn up*) consists of only 1 outcome. Hence $P(A^c) = 1/32$, and the answer is $P(A) = 1 - P(A^c) = 31/32$. ■

The next theorem is a simple extension of Axiom 3, which you can readily prove by induction.

THEOREM 2 **Addition Rule for Mutually Exclusive Events**

For mutually exclusive events $A_1, \cdots, A_m$ in a sample space S,

$$P(A_1 \cup A_2 \cup \cdots A_m) = P(A_1) + P(A_2) + \cdots + P(A_m). \tag{8}$$

EXAMPLE 3 **Mutually Exclusive Events**

If the probability that on any workday a garage will get 10–20, 21–30, 31–40, over 40 cars to service is 0.20, 0.35, 0.25, 0.12, respectively, what is the probability that on a given workday the garage gets at least 21 cars to service?

Solution. Since these are mutually exclusive events, Theorem 2 gives the answer $0.35 + 0.25 + 0.12 = 0.72$. Check this by the complementation rule. ■

In many cases, events will not be mutually exclusive. Then we have

THEOREM 3 **Addition Rule for Arbitrary Events**

For events A and B in a sample space,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B). \tag{9}$$

PROOF C, D, E in Fig. 512 make up $A \cup B$ and are mutually exclusive (disjoint). Hence by Theorem 2,

$$P(A \cup B) = P(C) + P(D) + P(E).$$

This gives (9) because on the right $P(C) + P(D) = P(A)$ by Axiom 3 and disjointness; and $P(E) = P(B) - P(D) = P(B) - P(A \cap B)$, also by Axiom 3 and disjointness. ■

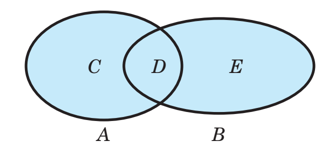


Fig. 512. Proof of Theorem 3

Note that for mutually exclusive events A and B we have $A \cap B = \varnothing$ by definition and, by comparing (9) and (6),

$$P(\varnothing) = 0. \tag{10}$$

(Can you also prove this by (5) and (7)?)

EXAMPLE 4 **Union of Arbitrary Events**

In tossing a fair die, what is the probability of getting an odd number or a number less than 4?

Solution. Let A be the event "*Odd number*" and B the event "*Number less than* 4." Then Theorem 3 gives the answer

$$P(A \cup B) = \tfrac{3}{6} + \tfrac{3}{6} - \tfrac{2}{6} = \tfrac{2}{3}$$

because $A \cap B =$ "*Odd number less than* 4" $= \{1, 3\}$. ■

Conditional Probability. Independent Events

Often it is required to find the probability of an event B under the condition that an event A occurs. This probability is called the **conditional probability** *of* B *given* A and is denoted by $P(B|A)$. In this case A serves as a new (reduced) sample space, and that probability is the fraction of $P(A)$ which corresponds to $A \cap B$. Thus

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) \neq 0]. \tag{11}$$

Similarly, the *conditional probability of* A *given* B is

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \qquad [P(B) \neq 0]. \tag{12}$$

Solving (11) and (12) for $P(A \cap B)$, we obtain

THEOREM 4 **Multiplication Rule**

If A and B are events in a sample space S and $P(A) \neq 0$, $P(B) \neq 0$, then

$$P(A \cap B) = P(A)P(B|A) = P(B)P(A|B). \tag{13}$$

EXAMPLE 5 **Multiplication Rule**

In producing screws, let A mean "screw too slim" and B "screw too short." Let $P(A) = 0.1$ and let the conditional probability that a slim screw is also too short be $P(B|A) = 0.2$. What is the probability that a screw that we pick randomly from the lot produced will be both too slim and too short?

Solution. $P(A \cap B) = P(A)P(B|A) = 0.1 \cdot 0.2 = 0.02 = 2\%$, by Theorem 4. ■

Independent Events. If events A and B are such that

$$P(A \cap B) = P(A)P(B), \tag{14}$$

they are called **independent events**. Assuming $P(A) \neq 0$, $P(B) \neq 0$, we see from (11)–(13) that in this case

$$P(A|B) = P(A), \qquad P(B|A) = P(B).$$

This means that the probability of A does not depend on the occurrence or nonoccurrence of B, and conversely. This justifies the term "independent."

Independence of *m* Events. Similarly, m events $A_1, \cdots, A_m$ are called **independent** if

$$P(A_1 \cap \cdots \cap A_m) = P(A_1) \cdots P(A_m) \tag{15a}$$

as well as for every k different events $A_{j_1}, A_{j_2}, \cdots, A_{j_k}$.

$$P(A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_k}) = P(A_{j_1})P(A_{j_2}) \cdots P(A_{j_k}) \tag{15b}$$

where $k = 2, 3, \cdots, m - 1$.

Accordingly, three events A, B, C are independent if and only if

$$\begin{aligned} P(A \cap B) &= P(A)P(B), \\ P(B \cap C) &= P(B)P(C), \\ P(C \cap A) &= P(C)P(A), \\ P(A \cap B \cap C) &= P(A)P(B)P(C). \end{aligned} \tag{16}$$

Sampling. Our next example has to do with randomly drawing objects, *one at a time,* from a given set of objects. This is called **sampling from a population**, and there are two ways of sampling, as follows.

1. In **sampling with replacement**, the object that was drawn at random is placed back to the given set and the set is mixed thoroughly. Then we draw the next object at random.
2. In **sampling without replacement** the object that was drawn is put aside.

EXAMPLE 6 Sampling With and Without Replacement

A box contains 10 screws, three of which are defective. Two screws are drawn at random. Find the probability that neither of the two screws is defective.

Solution. We consider the events

A: First drawn screw nondefective.

B: Second drawn screw nondefective.

Clearly, $P(A) = \frac{7}{10}$ because 7 of the 10 screws are nondefective and we sample at random, so that each screw has the same probability $(\frac{1}{10})$ of being picked. If we sample with replacement, the situation before the second drawing is the same as at the beginning, and $P(B) = \frac{7}{10}$. The events are independent, and the answer is

$$P(A \cap B) = P(A)P(B) = 0.7 \cdot 0.7 = 0.49 = 49\%.$$

If we sample without replacement, then $P(A) = \frac{7}{10}$, as before. If A has occurred, then there are 9 screws left in the box, 3 of which are defective. Thus $P(B|A) = \frac{6}{9} = \frac{2}{3}$, and Theorem 4 yields the answer

$$P(A \cap B) = \tfrac{7}{10} \cdot \tfrac{2}{3} = 47\%.$$

Is it intuitively clear that this value must be smaller than the preceding one? ■

PROBLEM SET 24.3

1. In rolling 3 fair dice, what is the probability of obtaining a sum not greater than 16?
2. In rolling 2 fair dice, what is the probability of a sum greater than 3 but not exceeding 6?
3. Three screws are drawn at random from a lot of 100 screws, 10 of which are defective. Find the probability of the event that all 3 screws drawn are nondefective, assuming that we draw **(a)** with replacement, **(b)** without replacement.
4. In Prob. 3 find the probability of *E: At least* 1 *defective* (i) directly, (ii) by using complements; in both cases **(a)** and **(b)**.
5. If a box contains 10 left-handed and 20 right-handed screws, what is the probability of obtaining at least one right-handed screw in drawing 2 screws with replacement?
6. Will the probability in Prob. 5 increase or decrease if we draw without replacement. First guess, then calculate.
7. Under what conditions will it make *practically* no difference whether we sample with or without replacement?
8. If a certain kind of tire has a life exceeding 40,000 miles with probability 0.90, what is the probability that a set of these tires on a car will last longer than 40,000 miles?
9. If we inspect photocopy paper by randomly drawing 5 sheets without replacement from every pack of 500, what is the probability of getting 5 clean sheets although 0.4% of the sheets contain spots?
10. Suppose that we draw cards repeatedly and with replacement from a file of 100 cards, 50 of which refer to male and 50 to female persons. What is the probability of obtaining the second "female" card before the third "male" card?
11. A batch of 200 iron rods consists of 50 oversized rods, 50 undersized rods, and 100 rods of the desired length. If two rods are drawn at random without replacement, what is the probability of obtaining **(a)** two rods of the desired length, **(b)** exactly one of the desired length, **(c)** none of the desired length?
12. If a circuit contains four automatic switches and we want that, with a probability of 99%, during a given time interval the switches to be all working, what probability of failure per time interval can we admit for a single switch?
13. A pressure control apparatus contains 3 electronic tubes. The apparatus will not work unless all tubes are operative. If the probability of failure of each tube during some interval of time is 0.04, what is the corresponding probability of failure of the apparatus?
14. Suppose that in a production of spark plugs the fraction of defective plugs has been constant at 2% over a long time and that this process is controlled every half hour by drawing and inspecting two just produced. Find the probabilities of getting **(a)** no defectives, **(b)** 1 defective, **(c)** 2 defectives. What is the sum of these probabilities?
15. What gives the greater probability of hitting at least once: **(a)** hitting with probability 1/2 and firing 1 shot, **(b)** hitting with probability 1/4 and firing 2 shots, **(c)** hitting with probability 1/8 and firing 4 shots? First guess.
16. You may wonder whether in (16) the last relation follows from the others, but the answer is no. To see this, imagine that a chip is drawn from a box containing 4 chips numbered 000, 011, 101, 110, and let A, B, C be the events that the first, second, and third digit, respectively, on the drawn chip is 1. Show that then the first three formulas in (16) hold but the last one does not hold.
17. Show that if B is a subset of A, then $P(B) \leqq P(A)$.
18. Extending Theorem 4, show that $P(A \cap B \cap C) = P(A)P(B|A)P(C|A \cap B)$.
19. Make up an example similar to Prob. 16, for instance, in terms of divisibility of numbers.

24.4 Permutations and Combinations

Permutations and combinations help in finding probabilities $P(A) = a/k$ by ***systematically counting*** the number a of points of which an event A consists; here, k is the number of points of the sample space S. The practical difficulty is that a may often be surprisingly large, so that actual counting becomes hopeless. For example, if in assembling some instrument you need 10 different screws in a certain order and you want to draw them

randomly from a box (which contains nothing else) the probability of obtaining them in the required order is only 1/3,628,800 because there are

$$10! = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 \cdot 6 \cdot 7 \cdot 8 \cdot 9 \cdot 10 = 3{,}628{,}800$$

orders in which they can be drawn. Similarly, in many other situations the numbers of orders, arrangements, etc. are often incredibly large. (If you are unimpressed, take 20 screws—how much bigger will the number be?)

Permutations

A **permutation** of given things (*elements* or *objects*) is an arrangement of these things in a row in some order. For example, for three letters a, b, c there are $3! = 1 \cdot 2 \cdot 3 = 6$ permutations: *abc*, *acb*, *bac*, *bca*, *cab*, *cba*. This illustrates (a) in the following theorem.

THEOREM 1

Permutations

(a) *Different things.* *The number of permutations of n different things taken all at a time is*

$$n! = 1 \cdot 2 \cdot 3 \cdots n \qquad \text{(read “}n\text{ factorial”)}. \tag{1}$$

(b) *Classes of equal things.* *If n given things can be divided into c classes of alike things differing from class to class, then the number of permutations of these things taken all at a time is*

$$\frac{n!}{n_1! n_2! \cdots n_c!} \qquad (n_1 + n_2 + \cdots + n_c = n) \tag{2}$$

Where n_j is the number of things in the jth class.

PROOF **(a)** There are n choices for filling the first place in the row. Then $n - 1$ things are still available for filling the second place, etc.

(b) n_1 alike things in class 1 make $n_1!$ permutations collapse into a single permutation (those in which class 1 things occupy the same n_1 positions), etc., so that (2) follows from (1). ■

EXAMPLE 1 **Illustration of Theorem 1(b)**

If a box contains 6 red and 4 blue balls, the probability of drawing first the red and then the blue balls is

$$P = 6!4!/10! = 1/210 \approx 0.5\%. \qquad ■$$

A **permutation of *n* things taken *k* at a time** is a permutation containing only k of the n given things. Two such permutations consisting of the same k elements, in a different order, are different, by definition. For example, there are 6 different permutations of the three letters a, b, c, taken two letters at a time, *ab*, *ac*, *bc*, *ba*, *ca*, *cb*.

A **permutation of *n* things taken *k* at a time with repetitions** is an arrangement obtained by putting any given thing in the first position, any given thing, including a repetition of the one just used, in the second, and continuing until k positions are filled. For example, there

are $3^2 = 9$ different such permutations of a, b, c taken 2 letters at a time, namely, the preceding 6 permutations and *aa*, *bb*, *cc*. You may prove (see Team Project 14):

THEOREM 2

Permutations

The number of different permutations of n different things taken k at a time without repetitions is

$$n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!} \tag{3a}$$

and with repetitions is

$$n^k. \tag{3b}$$

EXAMPLE 2 **Illustration of Theorem 2**

In an encrypted message the letters are arranged in groups of five letters, called *words*. From (3b) we see that the number of different such words is

$$26^5 = 11{,}881{,}376.$$

From (3a) it follows that the number of different such words containing each letter no more than once is

$$26!/(26-5)! = 26 \cdot 25 \cdot 24 \cdot 23 \cdot 22 = 7{,}893{,}600.$$ ■

Combinations

In a permutation, the order of the selected things is essential. In contrast, a **combination** of given things means any selection of one or more things *without regard to order*. There are two kinds of combinations, as follows.

The number of **combinations of *n* different things, taken *k* at a time, without repetitions** is the number of sets that can be made up from the n given things, each set containing k different things and no two sets containing exactly the same k things.

The number of **combinations of *n* different things, taken *k* at a time, with repetitions** is the number of sets that can be made up of k things chosen from the given n things, each being used as often as desired.

For example, there are three combinations of the three letters a, b, c, taken two letters at a time, without repetitions, namely, *ab*, *ac*, *bc*, and six such combinations with repetitions, namely, *ab*, *ac*, *bc*, *aa*, *bb*, *cc*.

THEOREM 3

Combinations

The number of different combinations of n different things taken, k at a time, without repetitions, is

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} = \frac{n(n-1)\cdots(n-k+1)}{1 \cdot 2 \cdots k}, \tag{4a}$$

and the number of those combinations with repetitions is

$$\binom{n+k-1}{k}. \tag{4b}$$

PROOF The statement involving (4a) follows from the first part of Theorem 2 by noting that there are $k!$ *permutations* of k things from the given n things that differ by the order of the elements (see Theorem 1), but there is only a single *combination* of those k things of the type characterized in the first statement of Theorem 3. The last statement of Theorem 3 can be proved by induction (see Team Project 14). ■

EXAMPLE 3 **Illustration of Theorem 3**

The number of samples of five lightbulbs that can be selected from a lot of 500 bulbs is [see (4a)]

$$\binom{500}{5} = \frac{500!}{5!495!} = \frac{500 \cdot 499 \cdot 498 \cdot 497 \cdot 496}{1 \cdot 2 \cdot 3 \cdot 4 \cdot 5} = 255{,}244{,}687{,}600. \quad ■$$

Factorial Function

In (1)–(4) the **factorial function** is basic. By definition,

$$0! = 1. \tag{5}$$

Values may be computed recursively from given values by

$$(n + 1)! = (n + 1)n!. \tag{6}$$

For large n the function is very large (see Table A3 in App. 5). A convenient approximation for large n is the **Stirling formula**[2]

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n \qquad (e = 2.718 \cdots) \tag{7}$$

where $\sim$ is read "**asymptotically equal**" and means that the ratio of the two sides of (7) approaches 1 as n approaches infinity.

EXAMPLE 4 **Stirling Formula**

$n!$	By (7)	Exact Value	Relative Error
4!	23.5	24	2.1%
10!	3,598,696	3,628,800	0.8%
20!	$2.42279 \cdot 10^{18}$	2,432,902,008,176,640,000	0.4%

■

Binomial Coefficients

The **binomial coefficients** are defined by the formula

$$\binom{a}{k} = \frac{a(a-1)(a-2)\cdots(a-k+1)}{k!} \qquad (k \geqq 0, \text{ integer}). \tag{8}$$

[2]JAMES STIRLING (1692–1770), Scots mathematician.

The numerator has k factors. Furthermore, we define

$$(9) \qquad \binom{a}{0} = 1, \qquad \text{in particular,} \qquad \binom{0}{0} = 1.$$

For integer $a = n$ we obtain from (8)

$$(10) \qquad \binom{n}{k} = \binom{n}{n-k} \qquad (n \geqq 0,\ 0 \leqq k \leqq n).$$

Binomial coefficients may be computed recursively, because

$$(11) \qquad \binom{a}{k} + \binom{a}{k+1} = \binom{a+1}{k+1} \qquad (k \geqq 0, \text{ integer}).$$

Formula (8) also yields

$$(12) \qquad \binom{-m}{k} = (-1)^k \binom{m+k-1}{k} \qquad \begin{matrix}(k \geqq 0, \text{ integer})\\ (m > 0).\end{matrix}$$

There are numerous further relations; we mention two important ones,

$$(13) \qquad \sum_{s=0}^{n-1} \binom{k+s}{k} = \binom{n+k}{k+1} \qquad \begin{matrix}(k \geqq 0,\ n \geqq 1,\\ \text{both integer})\end{matrix}$$

and

$$(14) \qquad \sum_{k=0}^{r} \binom{p}{k}\binom{q}{r-k} = \binom{p+q}{r} \qquad (r \geqq 0, \text{ integer}).$$

PROBLEM SET 24.4

Note the large numbers in the answers to some of these problems, which would make ***counting cases hopeless!***

1. In how many ways can a company assign 10 drivers to n buses, one driver to each bus and conversely?

2. List **(a)** all permutations, **(b)** all combinations without repetitions, **(c)** all combinations with repetitions, of 5 letters a, e, i, o, u taken 2 at a time.

3. If a box contains 4 rubber gaskets and 2 plastic gaskets, what is the probability of drawing **(a)** first the plastic and then the rubber gaskets, **(b)** first the rubber and then the plastic ones? Do this by using a theorem and checking it by multiplying probabilities.

4. An urn contains 2 green, 3 yellow, and 5 red balls. We draw 1 ball at random and put it aside. Then we draw the next ball, and so on. Find the probability of drawing at first the 2 green balls, then the 3 yellow ones, and finally the red ones.

5. In how many different ways can we select a committee consisting of 3 engineers, 2 physicists, and 2 computer scientists from 10 engineers, 5 physicists, and 6 computer scientists? First guess.

6. How many different samples of 4 objects can we draw from a lot of 50?

7. Of a lot of 10 items, 2 are defective. **(a)** Find the number of different samples of 4. Find the number of samples of 4 containing **(b)** no defectives, **(c)** 1 defective, **(d)** 2 defectives.

8. Determine the number of different bridge hands. (A bridge hand consists of 13 cards selected from a full deck of 52 cards.)

9. In how many different ways can 6 people be seated at a round table?

10. If a cage contains 100 mice, 3 of which are male, what is the probability that the 3 male mice will be included if 10 mice are randomly selected?

11. How many automobile registrations may the police have to check in a hit-and-run accident if a witness reports KDP7 and cannot remember the last two digits on the license plate but is certain that all three digits were different?

12. If 3 suspects who committed a burglary and 6 innocent persons are lined up, what is the probability that a witness who is not sure and has to pick three persons will pick the three suspects by chance? That the witness picks 3 innocent persons by chance?

13. CAS PROJECT. Stirling formula. (a) Using (7), compute approximate values of $n!$ for $n = 1, \cdots, 20$.

(b) Determine the relative error in (a). Find an empirical formula for that relative error.

(c) An upper bound for that relative error is $e^{1/12n} - 1$. Try to relate your empirical formula to this.

(d) Search through the literature for further information on Stirling's formula. Write a short eassy about your findings, arranged in logical order and illustrated with numeric examples.

14. TEAM PROJECT. Permutations, Combinations.

(a) Prove Theorem 2.

(b) Prove the last statement of Theorem 3.

(c) Derive (11) from (8).

(d) By the **binomial theorem**,

$$(a + b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k},$$

so that $a^k b^{n-k}$ has the coefficient $\binom{n}{k}$. Can you conclude this from Theorem 3 or is this a mere coincidence?

(e) Prove (14) by using the binomial theorem.

(f) Collect further formulas for binomial coefficients from the literature and illustrate them numerically.

15. Birthday problem. What is the probability that in a group of 20 people (that includes no twins) at least two have the same birthday, if we assume that the probability of having birthday on a given day is 1/365 for every day. First guess. *Hint.* Consider the complementary event.

24.5 Random Variables. Probability Distributions

In Sec. 24.1 we considered frequency distributions of data. These distributions show the absolute or relative frequency of the data values. Similarly, a **probability distribution** or, briefly, a **distribution**, shows the probabilities of events in an experiment. The quantity that we observe in an experiment will be denoted by X and called a **random variable** (or **stochastic variable**) because the value it will assume in the next trial depends on chance, on **randomness**—if you roll a die, you get one of the numbers from 1 to 6, but you don't know which one will show up next. Thus X = *Number a die turns up* is a random variable. So is X = *Elasticity of rubber* (elongation at break). ("Stochastic" means related to chance.)

If we ***count*** (cars on a road, defective screws in a production, tosses until a die shows the first Six), we have a **discrete random variable and distribution**. If we ***measure*** (electric voltage, rainfall, hardness of steel), we have a **continuous random variable and distribution**. Precise definitions follow. In both cases the distribution of X is determined by the **distribution function**

(1) $$F(x) = P(X \leqq x);$$

this is the probability that in a trial, X will assume any value not exceeding x.

CAUTION! The terminology is not uniform. $F(x)$ is sometimes also called the **cumulative distribution function**.

For (1) to make sense in both the discrete and the continuous case we formulate conditions as follows.

DEFINITION

Random Variable

A **random variable** X is a function defined on the sample space S of an experiment. Its values are real numbers. For every number a the probability

$$P(X = a)$$

with which X assumes a is defined. Similarly, for any interval I the probability

$$P(X \in I)$$

with which X assumes any value in I is defined.

Although this definition is very general, in practice only a very small number of distributions will occur over and over again in applications.

From (1) we obtain the fundamental formula for the probability corresponding to an interval $a < x \leqq b$,

$$P(a < X \leqq b) = F(b) - F(a). \tag{2}$$

This follows because $X \leqq a$ ("*X assumes any value not exceeding a*") and $a < X \leqq b$ ("*X assumes any value in the interval* $a < x \leqq b$") are mutually exclusive events, so that by (1) and Axiom 3 of Definition 2 in Sec. 24.3

$$\begin{aligned} F(b) = P(X \leqq b) &= P(X \leqq a) + P(a < X \leqq b) \\ &= F(a) + P(a < X \leqq b) \end{aligned}$$

and subtraction of $F(a)$ on both sides gives (2).

Discrete Random Variables and Distributions

By definition, a random variable X and its distribution are **discrete** if X assumes only finitely many or at most countably many values $x_1, x_2, x_3, \cdots$, called the **possible values** of X, with positive probabilities $p_1 = P(X = x_1), p_2 = P(X = x_2), \quad p_3 = P(X = x_3), \cdots$, whereas the probability $P(X \in I)$ is zero for any interval I containing no possible value.

Clearly, the discrete distribution of X is also determined by the **probability function** $f(x)$ of X, defined by

$$f(x) = \begin{cases} p_j & \text{if } x = x_j \\ 0 & \text{otherwise} \end{cases} \qquad (j = 1, 2, \cdots), \tag{3}$$

From this we get the values of the **distribution function** $F(x)$ by taking sums,

$$F(x) = \sum_{x_j \leqq x} f(x_j) = \sum_{x_j \leqq x} p_j \tag{4}$$

where for any given x we sum all the probabilities p_j for which x_j is smaller than or equal to that of x. This is a **step function** with upward jumps of size p_j at the possible values x_j of X and constant in between.

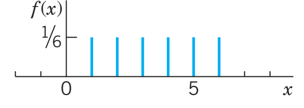
EXAMPLE 1 Probability Function and Distribution Function

Figure 513 shows the probability function $f(x)$ and the distribution function $F(x)$ of the discrete random variable

$$X = \textit{Number a fair die turns up}.$$

X has the possible values $x = 1, 2, 3, 4, 5, 6$ with probability $1/6$ each. At these x the distribution function has upward jumps of magnitude $1/6$. Hence from the graph of $f(x)$ we can construct the graph of $F(x)$ and conversely.

In Figure 513 (and the next one) at each jump the ***fat dot*** indicates the ***function value at the jump!*** ■



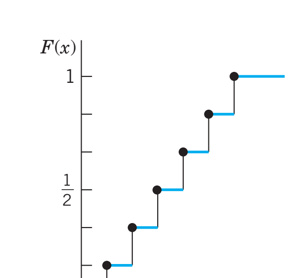
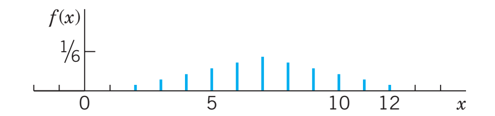


Fig. 513. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Number obtained in tossing a fair die once*



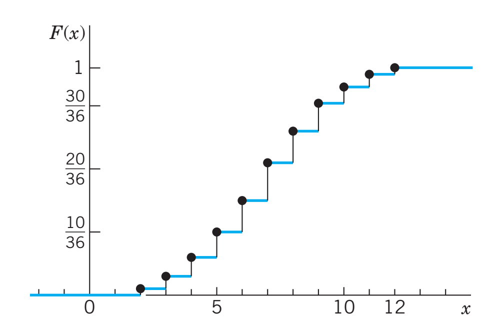


Fig. 514. Probability function $f(x)$ and distribution function $F(x)$ of the random variable X = *Sum of the two numbers obtained in tossing two fair dice once*

EXAMPLE 2 Probability Function and Distribution Function

The random variable X = *Sum of the two numbers two fair dice turn up* is discrete and has the possible values $2\,(= 1 + 1), 3, 4, \cdots, 12\,(= 6 + 6)$. There are $6 \cdot 6 = 36$ equally likely outcomes $(1, 1)\,(1, 2), \cdots, (6, 6)$, where the first number is that shown on the first die and the second number that on the other die. Each such outcome has probability $1/36$. Now $X = 2$ occurs in the case of the outcome $(1, 1)$; $X = 3$ in the case of the two outcomes $(1, 2)$ and $(2, 1)$; $X = 4$ in the case of the three outcomes $(1, 3), (2, 2), (3, 1)$; and so on. Hence $f(x) = P(X = x)$ and $F(x) = P(X \leqq x)$ have the values

x	2	3	4	5	6	7	8	9	10	11	12
$f(x)$	1/36	2/36	3/36	4/36	5/36	6/36	5/36	4/36	3/36	2/36	1/36
$F(x)$	1/36	3/36	6/36	10/36	15/36	21/36	26/36	30/36	33/36	35/36	36/36

Figure 514 shows a bar chart of this function and the graph of the distribution function, which is again a step function, with jumps (of different height!) at the possible values of X. ■

Two useful formulas for discrete distributions are readily obtained as follows. For the probability corresponding to intervals we have from (2) and (4)

$$P(a < X \leqq b) = F(b) - F(a) = \sum_{a<x_j \leqq b} p_j \qquad (X \text{ discrete}). \tag{5}$$

This is the sum of all probabilities p_j for which x_j satisfies $a < x_j \leqq b$. (Be careful about $<$ and $\leqq$!) From this and $P(S) = 1$ (Sec. 24.3) we obtain the following formula.

$$\sum_j p_j = 1 \qquad \text{(sum of all probabilities).} \tag{6}$$

EXAMPLE 3 Illustration of Formula (5)

In Example 2, compute the probability of a sum of at least 4 and at most 8.

Solution. $P(3 < X \leqq 8) = F(8) - F(3) = \frac{26}{36} - \frac{3}{36} = \frac{23}{36}$. ■

EXAMPLE 4 Waiting Time Problem. Countably Infinite Sample Space

In tossing a fair coin, let X = *Number of trials until the first head appears.* Then, by independence of events (Sec. 24.3),

$$\begin{aligned} P(X = 1) &= P(H) &&= \tfrac{1}{2} && && (H = \text{Head}) \\ P(X = 2) &= P(TH) &&= \tfrac{1}{2} \cdot \tfrac{1}{2} &&= \tfrac{1}{4} && (T = \text{Tail}) \\ P(X = 3) &= P(TTH) &&= \tfrac{1}{2} \cdot \tfrac{1}{2} \cdot \tfrac{1}{2} &&= \tfrac{1}{8}, && \text{etc.} \end{aligned}$$

and in general $P(X = n) = (\frac{1}{2})^n$, $n = 1, 2, \cdots$. Also, (6) can be confirmed by the sum formula for the geometric series,

$$\begin{aligned} \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots &= -1 + \frac{1}{1 - \frac{1}{2}} \\ &= -1 + 2 = 1. \end{aligned}$$ ■

Continuous Random Variables and Distributions

Discrete random variables appear in experiments in which we ***count*** (defectives in a production, days of sunshine in Chicago, customers standing in a line, etc.). Continuous random variables appear in experiments in which we ***measure*** (lengths of screws, voltage in a power line, Brinell hardness of steel, etc.). By definition, a random variable X and its distribution are *of continuous type* or, briefly, **continuous**, if its distribution function $F(x)$ [defined in (1)] can be given by an integral

$$F(x) = \int_{-\infty}^{x} f(v)\, dv \tag{7}$$

(we write v because x is needed as the upper limit of the integral) whose integrand $f(x)$, called the **density** of the distribution, is nonnegative, and is continuous, perhaps except for finitely many x-values. Differentiation gives the relation of f to F as

$$f(x) = F'(x) \tag{8}$$

for every x at which $f(x)$ is continuous.

From (2) and (7) we obtain the very important formula for the probability corresponding to an interval:

$$P(a < X \leqq b) = F(b) - F(a) = \int_a^b f(v)\, dv. \tag{9}$$

This is the analog of (5).

From (7) and $P(S) = 1$ (Sec. 24.3) we also have the analog of (6):

$$\int_{-\infty}^{\infty} f(v)\, dv = 1. \tag{10}$$

Continuous random variables are ***simpler than discrete ones*** with respect to intervals. Indeed, in the continuous case the four probabilities corresponding to $a < X \leqq b$, $a < X < b$, $a \leqq X < b$, and $a \leqq X \leqq b$ with any fixed a and b $(> a)$ are all the same. Can you see why? (*Answer.* This probability is the area under the density curve, as in Fig. 515, and does not change by adding or subtracting a single point in the interval of integration.) This is different from the discrete case! (Explain.)

The next example illustrates notations and typical applications of our present formulas.

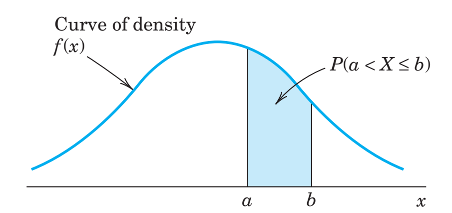


Fig. 515. Example illustrating formula (9)

EXAMPLE 5 Continuous Distribution

Let X have the density function $f(x) = 0.75(1 - x^2)$ if $-1 \leqq x \leqq 1$ and zero otherwise. Find the distribution function. Find the probabilities $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2})$ and $P(\frac{1}{4} \leqq X \leqq 2)$. Find x such that $P(X \leqq x) = 0.95$.

Solution. From (7) we obtain $F(x) = 0$ if $x \leqq -1$,

$$F(x) = 0.75 \int_{-1}^{x} (1 - v^2)\, dv = 0.5 + 0.75x - 0.25x^3 \qquad \text{if } -1 < x \leqq 1,$$

and $F(x) = 1$ if $x > 1$. From this and (9) we get

$$P(-\tfrac{1}{2} \leqq X \leqq \tfrac{1}{2}) = F(\tfrac{1}{2}) - F(-\tfrac{1}{2}) = 0.75 \int_{-1/2}^{1/2} (1 - v^2)\, dv = 68.75\%$$

(because $P(-\frac{1}{2} \leqq X \leqq \frac{1}{2}) = P(-\frac{1}{2} < X \leqq \frac{1}{2})$ for a continuous distribution) and

$$P(\tfrac{1}{4} \leqq X \leqq 2) = F(2) - F(\tfrac{1}{4}) = 0.75 \int_{1/4}^{1} (1 - v^2)\, dv = 31.64\%.$$

(Note that the upper limit of integration is 1, not 2. Why?) Finally,

$$P(X \leqq x) = F(x) = 0.5 + 0.75x - 0.25x^3 = 0.95.$$

Algebraic simplification gives $3x - x^3 = 1.8$. A solution is $x = 0.73$, approximately.

Sketch $f(x)$ and mark $x = -\frac{1}{2}, \frac{1}{2}, \frac{1}{4}$, and 0.73, so that you can see the results (the probabilities) as areas under the curve. Sketch also $F(x)$. ■

Further examples of continuous distributions are included in the next problem set and in later sections.

PROBLEM SET 24.5

1. Graph the probability function $f(x) = kx^2$ ($x = 1, 2, 3, 4, 5$; k suitable) and the distribution function.

2. Graph the density function $f(x) = kx^2$ ($0 \leqq x \leqq 5$; k suitable) and the distribution function.

3. **Uniform distribution.** Graph f and F when the density of X is $f(x) = k = \text{const}$ if $-2 \leqq x \leqq 2$ and 0 elsewhere. Find $P(0 \leqq X \leqq 2)$.

4. In Prob. 3 find c and $\tilde{c}$ such that $P(-c < X < c) = 95\%$ and $P(0 < X < \tilde{c}) = 95\%$.

5. Graph f and F when $f(-2) = f(2) = \frac{1}{8}$, $f(-1) = f(1) = \frac{3}{8}$. Can f have further positive values?

6. A box contains 4 right-handed and 6 left-handed screws. Two screws are drawn at random without replacement. Let X be the number of left-handed screws drawn. Find the probabilities $P(X = 0)$, $P(X = 1)$, $P(X = 2)$, $P(1 < X < 2)$, $P(X \leqq 1)$, $P(X \geqq 1)$, $P(X > 1)$, and $P(0.5 < X < 10)$.

7. Let X be the number of years before a certain kind of pump needs replacement. Let X have the probability function $f(x) = kx^3$, $x = 0, 1, 2, 3, 4$, Find k. Sketch f and F.

8. Graph the distribution function $F(x) = 1 - e^{-3x}$ if $x > 0$, $F(x) = 0$ if $x \leqq 0$, and the density $f(x)$. Find x such that $F(x) = 0.9$.

9. Let X [millimeters] be the thickness of washers. Assume that X has the density $f(x) = kx$ if $0.9 < x < 1.1$ and 0 otherwise. Find k. What is the probability that a washer will have thickness between 0.95 mm and 1.05 mm?

10. If the diameter X of axles has the density $f(x) = k$ if $119.9 \leqq x \leqq 120.1$ and 0 otherwise, how many defectives will a lot of 500 axles approximately contain if defectives are axles slimmer than 119.91 or thicker than 120.09?

11. Find the probability that none of three bulbs in a traffic signal will have to be replaced during the first 1500 hours of operation if the lifetime X of a bulb is a random variable with the density $f(x) = 6[0.25 - (x - 1.5)^2]$ when $1 \leqq x \leqq 2$ and $f(x) = 0$ otherwise, where x is measured in multiples of 1000 hours.

12 Let X be the ratio of sales to profits of some company. Assume that X has the distribution function $F(x) = 0$ if $x < 2$, $F(x) = (x^2 - 4)/5$ if $2 \leqq x < 3$, $F(x) = 1$ if $x \geqq 3$. Find and sketch the density. What is the probability that X is between 2.5 (40% profit) and 5 (20% profit)?

13. Suppose that in an automatic process of filling oil cans, the content of a can (in gallons) is $Y = 100 + X$, where X is a random variable with density $f(x) = 1 - |x|$ when $|x| \leqq 1$ and 0 when $|x| > 1$. Sketch $f(x)$ and $F(x)$. In a lot of 1000 cans, about how many will contain 100 gallons or more? What is the probability that a can will contain less than 99.5 gallons? Less than 99 gallons?

14. Find the probability function of X = *Number of times a fair die is rolled until the first Six appears* and show that it satisfies (6).

15. Let X be a random variable that can assume every real value. What are the complements of the events $X \leqq b$, $X < b$, $X \geqq c$, $X > c$, $b \leqq X \leqq c$, $b < X \leqq c$?

24.6 Mean and Variance of a Distribution

The mean μ and variance σ^2 of a random variable X and of its distribution are the theoretical counterparts of the mean $\bar{x}$ and variance s^2 of a frequency distribution in Sec. 24.1 and serve a similar purpose. Indeed, the mean characterizes the central location and the variance the spread (the variability) of the distribution. The **mean** μ (mu) is defined by

(1)

$$\textbf{(a)} \quad \mu = \sum_j x_j f(x_j) \qquad \text{(Discrete distribution)}$$

$$\textbf{(b)} \quad \mu = \int_{-\infty}^{\infty} x f(x)\, dx \qquad \text{(Continuous distribution)}$$

and the **variance** σ^2 (sigma square) by

(2)

$$\textbf{(a)} \quad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \qquad \text{(Discrete distribution)}$$

$$\textbf{(b)} \quad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\, dx \qquad \text{(Continuous distribution).}$$

σ (the positive square root of σ^2) is called the **standard deviation** of X and its distribution. f is the probability function or the density, respectively, in (a) and (b).

The mean μ is also denoted by $E(X)$ and is called the **expectation** *of* X because it gives the average value of X to be expected in many trials. Quantities such as μ and σ^2 that measure certain properties of a distribution are called **parameters**. μ and σ^2 are the two most important ones. From (2) we see that

$$\sigma^2 > 0 \tag{3}$$

(except for a discrete "distribution" with only one possible value, so that $\sigma^2 = 0$). We assume that μ and σ^2 exist (are finite), as is the case for practically all distributions that are useful in applications.

EXAMPLE 1 Mean and Variance

The random variable X = *Number of heads in a single toss of a fair coin* has the possible values $X = 0$ and $X = 1$ with probabilities $P(X = 0) = \frac{1}{2}$ and $P(X = 1) = \frac{1}{2}$. From (la) we thus obtain the mean $\mu = 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}$, and (2a) yields the variance

$$\sigma^2 = (0 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} + (1 - \tfrac{1}{2})^2 \cdot \tfrac{1}{2} = \tfrac{1}{4}.$$ ■

EXAMPLE 2 Uniform Distribution. Variance Measures Spread

The distribution with the density

$$f(x) = \frac{1}{b - a} \qquad \text{if} \qquad a < x < b$$

and $f = 0$ otherwise is called the **uniform distribution** on the interval $a < x < b$. From (1b) (or from Theorem 1, below) we find that $\mu = (a + b)/2$, and (2b) yields the variance

$$\sigma^2 = \int_a^b \left(x - \frac{a+b}{2}\right)^2 \frac{1}{b-a}\, dx = \frac{(b-a)^2}{12}.$$

Figure 516 illustrates that the spread is large if and only if σ^2 is large. ■

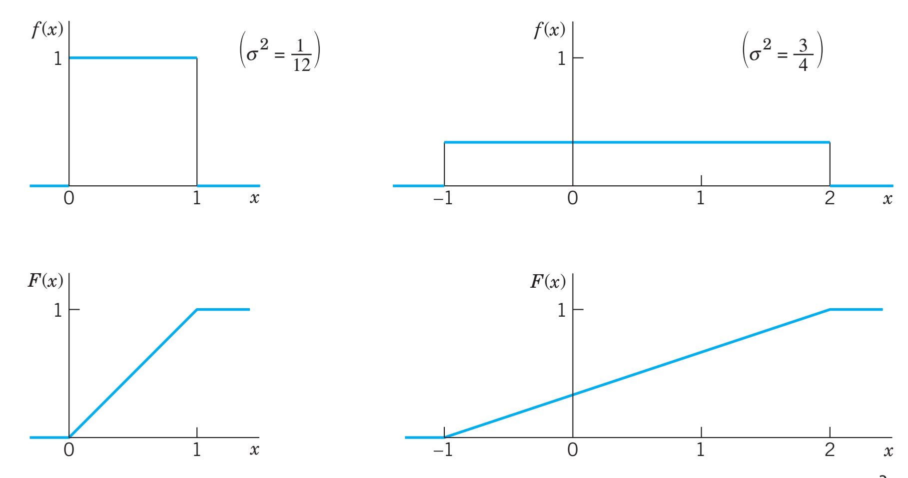


Fig. 516. Uniform distributions having the same mean (0.5) but different variances σ^2

Symmetry. We can obtain the mean μ without calculation if a distribution is symmetric. Indeed, you may prove

THEOREM 1

Mean of a Symmetric Distribution

If a distribution is **symmetric** *with respect to* $x = c$, *that is,* $f(c - x) = f(c + x)$, *then* $\mu = c$. (Examples 1 and 2 illustrate this.)

Transformation of Mean and Variance

Given a random variable X with mean μ and variance σ^2, we want to calculate the mean and variance of $X^* = a_1 + a_2X$, where a_1 and a_2 are given constants. This problem is important in statistics, where it often appears.

THEOREM 2

Transformation of Mean and Variance

(a) *If a random variable X has mean* μ *and variance* σ^2, *then the random variable*

$$X^* = a_1 + a_2X \qquad (a_2 > 0) \tag{4}$$

has the mean μ^* *and variance* σ^{*2}, *where*

$$\mu^* = a_1 + a_2\mu \qquad \text{and} \qquad \sigma^{*2} = a_2^2\sigma^2. \tag{5}$$

(b) *In particular, the* **standardized random variable** *Z corresponding to X, given by*

(6)
$$Z = \frac{X - \mu}{\sigma}$$

has the mean 0 *and the variance* 1.

PROOF We prove (5) for a continuous distribution. To a small interval I of length Δx on the x-axis there corresponds the probability $f(x)\Delta x$ [approximately; the area of a rectangle of base Δx and height $f(x)$]. Then the probability $f(x)\Delta x$ must equal that for the corresponding interval on the x^*-axis, that is, $f^*(x^*)\Delta x^*$, where f^* is the density of X^* and Δx^* is the length of the interval on the x^*-axis corresponding to I. Hence for differentials we have $f^*(x^*)\,dx^* = f(x)\,dx$. Also, $x^* = a_1 + a_2 x$ by (4), so that (1b) applied to X^* gives

$$\begin{aligned}\mu^* &= \int_{-\infty}^{\infty} x^* f^*(x^*)\,dx^* \\ &= \int_{-\infty}^{\infty} (a_1 + a_2 x) f(x)\,dx \\ &= a_1 \int_{-\infty}^{\infty} f(x)\,dx + a_2 \int_{-\infty}^{\infty} x f(x)\,dx.\end{aligned}$$

On the right the first integral equals 1, by (10) in Sec. 24.5. The second intergral is μ. This proves (5) for μ^*. It implies

$$x^* - \mu^* = (a_1 + a_2 x) - (a_1 + a_2\mu) = a_2(x - \mu).$$

From this and (2) applied to X^*, again using $f^*(x^*)\,dx^* = f(x)\,dx$, we obtain the second formula in (5),

$$\sigma^{*2} = \int_{-\infty}^{\infty} (x^* - \mu^*)^2 f^*(x^*)\,dx^* = a_2^2 \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx = a_2^2\sigma^2.$$

For a discrete distribution the proof of (5) is similar.

Choosing $a_1 = -\mu/\sigma$ and $a_2 = 1/\sigma$ we obtain (6) from (4), writing $X^* = Z$. For these a_1, a_2 formula (5) gives $\mu^* = 0$ and $\sigma^{*2} = 1$, as claimed in (b). ∎

Expectation, Moments

Recall that (1) defines the expectation (the mean) of X, the value of X to be expected on the average, written $\mu = E(X)$. More generally, if $g(x)$ is nonconstant and continuous for all x, then $g(X)$ is a random variable. Hence its *mathematical expectation* or, briefly, its

expectation $E(g(X))$ is the value of $g(X)$ to be expected on the average, defined [similarly to (1)] by

$$(7) \qquad E(g(X)) = \sum_j g(x_j)f(x_j) \qquad \text{or} \qquad E(g(X)) = \int_{-\infty}^{\infty} g(x)f(x)\,dx.$$

In the first formula, f is the probability function of the discrete random variable X. In the second formula, f is the density of the continuous random variable X. Important special cases are the ***k*th moment** of X (where $k = 1, 2, \cdots$)

$$(8) \qquad E(X^k) = \sum_j x_j^k f(x_j) \qquad \text{or} \qquad \int_{-\infty}^{\infty} x^k f(x)\,dx$$

and the ***k*th central moment** of X ($k = 1, 2, \cdots$)

$$(9) \qquad E([X-\mu]^k) = \sum_j (x_j - \mu)^k f(x_j) \qquad \text{or} \qquad \int_{-\infty}^{\infty} (x-\mu)^k f(x)\,dx.$$

This includes the first moment, the **mean** of X

$$\textbf{(10)} \qquad \mu = E(X) \qquad [(8) \text{ with } k = 1].$$

It also includes the second central moment, the **variance** of X

$$\textbf{(11)} \qquad \sigma^2 = E([X-\mu]^2) \qquad [(9) \text{ with } k = 2].$$

For later use you may prove

$$(12) \qquad E(1) = 1.$$

PROBLEM SET 24.6

1–8 MEAN, VARIANCE

Find the mean and variance of the random variable X with probability function or density $f(x)$.

1. $f(x) = kx$ ($0 \leqq x \leqq 2$, k suitable)
2. X = Number a fair die turns up
3. Uniform distribution on $[0, 2\pi]$
4. $Y = \sqrt{3}(X - \mu)/\pi$ with X as in Prob. 3
5. $f(x) = 4e^{-4x}$ ($x \geqq 0$)
6. $f(x) = k(1 - x^2)$ if $-1 \leqq x \leqq 1$ and 0 otherwise
7. $f(x) = Ce^{-x/2}$ $(x = 0)$
8. X = *Number of times a fair coin is flipped until the first Head appears.* (Calculate μ only.)
9. If the diameter X [cm] of certain bolts has the density $f(x) = k(x - 0.9)(1.1 - x)$ for $0.9 < x < 1.1$ and 0 for other x, what are k, μ, and σ^2? Sketch $f(x)$.
10. If, in Prob. 9, a defective bolt is one that deviates from 1.00 cm by more than 0.06 cm, what percentage of defectives should we expect?
11. For what choice of the maximum possible deviation from 1.00 cm shall we obtain 10% defectives in Probs. 9 and 10?
12. What total sum can you expect in rolling a fair die 20 times? Do the experiment. Repeat it a number of times and record how the sum varies.
13. What is the expected daily profit if a store sells X air conditioners per day with probability $f(10) = 0.1$, $f(11) = 0.3$, $f(12) = 0.4$, $f(13) = 0.2$ and the profit per conditioner is \$55?
14. Find the expectation of $g(X) = X^2$, where X is uniformly distributed on the interval $-1 \leqq x \leqq 1$.

15. A small filling station is supplied with gasoline every **Saturday** afternoon. Assume that its volume X of sales in ten thousands of gallons has the probability density $f(x) = 6x(1 - x)$ if $0 \leqq x \leqq 1$ and 0 otherwise. Determine the mean, the variance, and the standardized variable.

16. What capacity must the tank in Prob. 15 have in order that the probability that the tank will be emptied in a given week be 5%?

17. James rolls 2 fair dice, and Harry pays k cents to James, where k is the product of the two faces that show on the dice. How much should James pay to Harry for each game to make the game fair?

18. What is the mean life of a lightbulb whose life X [hours] has the density $f(x) = 0.001e^{-0.001x}$ $(x \geqq 0)$?

19. Let X be discrete with probability function $f(0) = f(3) = \frac{1}{8}$, $f(1) = f(2) = \frac{3}{8}$. Find the expectation of X^3.

20. TEAM PROJECT. Means, Variances, Expectations.

(a) Show that $E(X - \mu) = 0$, $\sigma^2 = E(X^2) - \mu^2$.

(b) Prove (10)–(12).

(c) Find all the moments of the uniform distribution on an interval $a \leqq x \leqq b$.

(d) The **skewness** γ of a random variable X is defined by

$$\gamma = \frac{1}{\sigma^3} E([X - \mu]^3). \tag{13}$$

Show that for a symmetric distribution (whose third central moment exists) the skewness is zero.

(e) Find the skewness of the distribution with density $f(x) = xe^{-x}$ when $x > 0$ and $f(x) = 0$ otherwise. Sketch $f(x)$.

(f) Calculate the skewness of a few simple discrete distributions of your own choice.

(g) Find a *nonsymmetric* discrete distribution with 3 possible values, mean 0, and skewness 0.

24.7 Binomial, Poisson, and Hypergeometric Distributions

These are the three most important ***discrete*** distributions, with numerous applications.

Binomial Distribution

The **binomial distribution** occurs in games of chance (rolling a die, see below, etc.), quality inspection (e.g., counting of the number of defectives), opinion polls (counting number of employees favoring certain schedule changes, etc.), medicine (e.g., recording the number of patients who recovered on a new medication), and so on. The conditions of its occurrence are as follows.

We are interested in the number of times an event A occurs in n independent trials. In each trial the event A has the same probability $P(A) = p$. Then in a trial, A will *not* occur with probability $q = 1 - p$. In n trials the random variable that interests us is

$$X = \textit{Number of times the event A occurs in n trials.}$$

X can assume the values $0, 1, \cdots, n$, and we want to determine the corresponding probabilities. Now $X = x$ means that A occurs in x trials and in $n - x$ trials it does not occur. This may look as follows.

$$\underbrace{A \quad A \cdots A}_{x \text{ times}} \qquad \underbrace{B \quad B \cdots B}_{n - x \text{ times}}. \tag{1}$$

Here $B = A^c$ is the complement of A, meaning that A does not occur (Sec. 24.2). We now use the assumption that the trials are independent, that is, they do not influence each other. Hence (1) has the probability (see Sec. 24.3 on independent events)

$$\underbrace{pp \cdots p}_{x \text{ times}} \cdot \underbrace{qq \cdots q}_{n - x \text{ times}} = p^x q^{n-x}. \tag{1*}$$

Now (1) is just one order of arranging x A's and $n - x$ B's. We now use Theorem 1(b) in Sec. 24.4, which gives the number of permutations of n things (the n outcomes of the n trials) consisting of 2 classes, class 1 containing the $n_1 = x$ A's and class 2 containing the $n - n_1 = n - x$ B's. This number is

$$\frac{n!}{x!(n-x)!} = \binom{n}{x}.$$

Accordingly, (1*), multiplied by this binomial coefficient, gives the probability $P(X = x)$ of $X = x$, that is, of obtaining A precisely x times in n trials. Hence X has the probability function

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n) \tag{2}$$

and $f(x) = 0$ otherwise. The distribution of X with probability function (2) is called the **binomial distribution** or *Bernoulli distribution*. The occurrence of A is called *success* (regardless of what it actually is; it may mean that you miss your plane or lose your watch) and the nonoccurrence of A is called *failure*. Figure 517 shows typical examples. Numeric values can be obtained from Table A5 in App. 5 or from your CAS.

The mean of the binomial distribution is (see Team Project 16)

$$\mu = np \tag{3}$$

and the variance is (see Team Project 16)

$$\sigma^2 = npq. \tag{4}$$

For the ***symmetric case*** of equal chance of success and failure ($p = q = \frac{1}{2}$) this gives the mean $n/2$, the variance $n/4$, and the probability function

$$f(x) = \binom{n}{x} \left(\frac{1}{2}\right)^n \qquad (x = 0, 1, \cdots, n). \tag{2*}$$

Fig. 517. Probability function (2) of the binomial distribution for $n = 5$ and various values of p

EXAMPLE 1 Binomial Distribution

Compute the probability of obtaining at least two *"Six"* in rolling a fair die 4 times.

Solution. $p = P(A) = P(\text{"Six"}) = \frac{1}{6}$, $q = \frac{5}{6}$, $n = 4$. The event *"At least two 'Six'"* occurs if we obtain 2 or 3 or 4 "Six." Hence the answer is

$$P = f(2) + f(3) + f(4) = \binom{4}{2}\left(\frac{1}{6}\right)^2\left(\frac{5}{6}\right)^2 + \binom{4}{3}\left(\frac{1}{6}\right)^3\left(\frac{5}{6}\right) + \binom{4}{4}\left(\frac{1}{6}\right)^4$$

$$= \frac{1}{6^4}(6 \cdot 25 + 4 \cdot 5 + 1) = \frac{171}{1296} = 13.2\%.$$

Poisson Distribution

The discrete distribution with infinitely many possible values and probability function

$$f(x) = \frac{\mu^x}{x!} e^{-\mu} \qquad (x = 0, 1, \cdots) \tag{5}$$

is called the **Poisson distribution**, named after S. D. Poisson (Sec. 18.5). Figure 518 shows (5) for some values of μ. It can be proved that this distribution is obtained as a limiting case of the binomial distribution, if we let $p \to 0$ and $n \to \infty$ so that the mean $\mu = np$ approaches a finite value. (For instance, $\mu = np$ may be kept constant.) The Poisson distribution has the mean μ and the variance (see Team Project 16)

$$\sigma^2 = \mu. \tag{6}$$

Figure 518 gives the impression that, with increasing mean, the spread of the distribution increases, thereby illustrating formula (6), and that the distribution becomes more and more (approximately) symmetric.

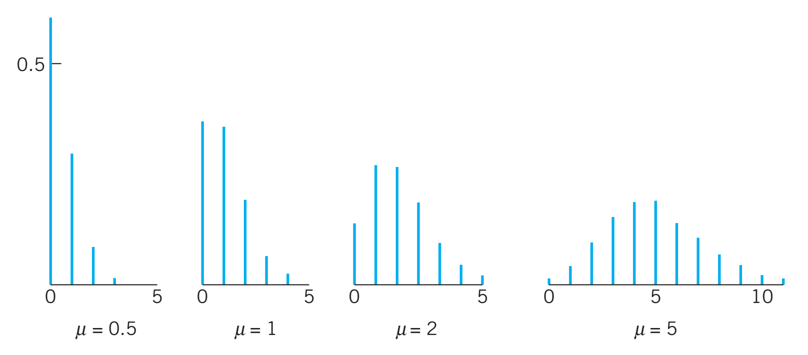


Fig. 518. Probability function (5) of the Poisson distribution for various values of μ

EXAMPLE 2 Poisson Distribution

If the probability of producing a defective screw is $p = 0.01$, what is the probability that a lot of 100 screws will contain more than 2 defectives?

Solution. The complementary event is A^c: *Not more than* 2 *defectives.* For its probability we get, from the binomial distribution with mean $\mu = np = 1$, the value [see (2)]

$$P(A^c) = \binom{100}{0} 0.99^{100} + \binom{100}{1} 0.01 \cdot 0.99^{99} + \binom{100}{2} 0.01^2 \cdot 0.99^{98}.$$

Since p is very small, we can approximate this by the much more convenient Poisson distribution with mean $\mu = np = 100 \cdot 0.01 = 1$, obtaining [see (5)]

$$P(A^c) \approx e^{-1}\,(1 + 1 + \tfrac{1}{2})$$
$$= 91.97\%.$$

Thus $P(A) = 8.03\%$. Show that the binomial distribution gives $P(A) = 7.94\%$, so that the Poisson approximation is quite good. ■

EXAMPLE 3 Parking Problems. Poisson Distribution

If on the average, 2 cars enter a certain parking lot per minute, what is the probability that during any given minute 4 or more cars will enter the lot?

Solution. To understand that the Poisson distribution is a model of the situation, we imagine the minute to be divided into very many short time intervals, let p be the (constant) probability that a car will enter the lot during any such short interval, and assume independence of the events that happen during those intervals. Then we are dealing with a binomial distribution with very large n and very small p, which we can approximate by the Poisson distribution with

$$\mu = np = 2,$$

because 2 cars enter on the average. The complementary event of the event "4 cars or more during a given minute" is "3 *cars or fewer enter the lot*" and has the probability

$$f(0) + f(1) + f(2) + f(3) = e^{-2}\left(\frac{2^0}{0!} + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!}\right)$$
$$= 0.857.$$

Answer: 14.3%. (Why did we consider that complement?) ■

Sampling with Replacement

This means that we draw things from a given set one by one, and after each trial we replace the thing drawn (put it back to the given set and mix) before we draw the next thing. This guarantees independence of trials and leads to the **binomial distribution**. Indeed, if a box contains N things, for example, screws, M of which are defective, the probability of drawing a defective screw in a trial is $p = M/N$. Hence the probability of drawing a nondefective screw is $q = 1 - p = 1 - M/N$, and (2) gives the probability of drawing x defectives in n trials in the form

$$f(x) = \binom{n}{x}\left(\frac{M}{N}\right)^x \left(1 - \frac{M}{N}\right)^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{7}$$

Sampling without Replacement. Hypergeometric Distribution

Sampling without replacement means that we return no screw to the box. Then we no longer have independence of trials (why?), and instead of (7) the probability of drawing x defectives in n trials is

(8)
$$f(x) = \frac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} \qquad (x = 0, 1, \cdots, n).$$

The distribution with this probability function is called the **hypergeometric distribution** (because its moment generating function (see Team Project 16) can be expressed by the hypergeometric function defined in Sec. 5.4, a fact that we shall not use).

Derivation of (8). By (4a) in Sec. 24.4 there are

(a) $\binom{N}{n}$ different ways of picking n things from N,

(b) $\binom{M}{x}$ different ways of picking x defectives from M,

(c) $\binom{N-M}{n-x}$ different ways of picking $n - x$ nondefectives from $N - M$,

and each way in (b) combined with each way in (c) gives the total number of mutually exclusive ways of obtaining x defectives in n drawings without replacement. Since (a) is the total number of outcomes and we draw at random, each such way has the probability $1\Big/\binom{N}{n}$. From this, (8) follows. ■

The hypergeometric distribution has the mean (Team Project 16)

(9)
$$\mu = n\frac{M}{N}$$

and the variance

(10)
$$\sigma^2 = \frac{nM(N-M)(N-n)}{N^2(N-1)}.$$

EXAMPLE 4 Sampling with and without Replacement

We want to draw random samples of two gaskets from a box containing 10 gaskets, three of which are defective. Find the probability function of the random variable X = *Number of defectives in the sample*.

Solution. We have $N = 10$, $M = 3$, $N - M = 7$, $n = 2$. For sampling with replacement, (7) yields

$$f(x) = \binom{2}{x}\left(\frac{3}{10}\right)^x\left(\frac{7}{10}\right)^{2-x}, \qquad f(0) = 0.49, \quad f(1) = 0.42, \quad f(2) = 0.09.$$

For sampling without replacement we have to use (8), finding

$$f(x) = \binom{3}{x}\binom{7}{2-x}\Big/\binom{10}{2}, \qquad f(0) = f(1) = \frac{21}{45} \approx 0.47, \quad f(2) = \frac{3}{45} \approx 0.07.$$ ■

If N, M, and N − M are large compared with n, then it does not matter too much whether we sample with or without replacement, and in this case the hypergeometric distribution may be approximated by the binomial distribution (with $p = M/N$), *which is somewhat simpler.*

Hence, in sampling from an indefinitely large population (“**infinite population**”), *we may use the binomial distribution, regardless of whether we sample with or without replacement.*

PROBLEM SET 24.7

1. Mark the positions of μ in Fig. 517. Comment.
2. Graph (2) for $n = 8$ as in Fig. 517 and compare with Fig. 517.
3. In Example 3, if 5 cars enter the lot on the average, what is the probability that during any given minute 6 or more cars will enter? First guess. Compare with Example 3.
4. How do the probabilities in Example 4 of the text change if you double the numbers: drawing 4 gaskets from 20, 6 of which are defective? First guess.
5. Five fair coins are tossed simultaneously. Find the probability function of the random variable X = *Number of heads* and compute the probabilities of obtaining no heads, precisely 1 head, at least 1 head, not more than 4 heads.
6. Suppose that 4% of steel rods made by a machine are defective, the defectives occurring at random during production. If the rods are packaged 100 per box, what is the Poisson approximation of the probability that a given box will contain $x = 0, 1, \cdots, 5$ defectives?
7. Let X be the number of cars per minute passing a certain point of some road between 8 A.M. and 10 A.M. on a Sunday. Assume that X has a Poisson distribution with mean 5. Find the probability of observing 4 or fewer cars during any given minute.
8. Suppose that a telephone switchboard of some company on the average handles 300 calls per hour, and that the board can make at most 10 connections per minute. Using the Poisson distribution, estimate the probability that the board will be overtaxed during a given minute. (Use Table A6 in App. 5 or your CAS.)
9. **Rutherford–Geiger experiments.** In 1910, E. Rutherford and H. Geiger showed experimentally that the number of alpha particles emitted per second in a radioactive process is a random variable X having a Poisson distribution. If X has mean 0.5, what is the probability of observing two or more particles during any given second?
10. Let $p = 2\%$ be the probability that a certain type of lightbulb will fail in a 24-hour test. Find the probability that a sign consisting of 15 such bulbs will burn 24 hours with no bulb failures.
11. Guess how much less the probability in Prob. 10 would be if the sign consisted of 100 bulbs. Then calculate.
12. Suppose that a certain type of magnetic tape contains, on the average, 2 defects per 100 meters. What is the probability that a roll of tape 300 meters long will contain **(a)** x defects, **(b)** no defects?
13. Suppose that a test for extrasensory perception consists of naming (in any order) 3 cards randomly drawn from a deck of 13 cards. Find the probability that by chance alone, the person will correctly name **(a)** no cards, **(b)** 1 card, **(c)** 2 cards, **(d)** 3 cards.
14. If a ticket office can serve at most 4 customers per minute and the average number of customers is 120 per hour, what is the probability that during a given minute customers will have to wait? (Use the Poisson distribution, Table 6 in Appendix 5.)
15. Suppose that in the production of 60-ohm radio resistors, nondefective items are those that have a resistance between 58 and 62 ohms and the probability of a resistor's being defective is 0.1%. The resistors are sold in lots of 200, with the guarantee that all resistors are nondefective. What is the probability that a given lot will violate this guarantee? (Use the Poisson distribution.)
16. **TEAM PROJECT. Moment Generating Function.** The moment generating function $G(t)$ is defined by

$$G(t) = E(e^{tX_j}) = \sum_j e^{tx_j} f(x_j)$$

or

$$G(t) = E(e^{tX}) = \int_{-\infty}^{\infty} e^{tx} f(x)\, dx$$

where X is a discrete or continuous random variable, respectively.

(a) Assuming that termwise differentiation and differentiation under the integral sign are permissible, show

that $E(X^k) = G^{(k)}(0)$, where $G^{(k)} = d^k G/dt^k$, in particular, $\mu = G'(0)$.

(b) Show that the binomial distribution has the moment generating function

$$G(t) = \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x q^{n-x}$$
$$= (pe^t + q)^n.$$

(c) Using (b), prove (3).

(d) Prove (4).

(e) Show that the Poisson distribution has the moment generating function $G(t) = e^{-\mu} e^{\mu e^t}$ and prove (6).

(f) Prove $x\binom{M}{x} = M\binom{M-1}{x-1}$.

Using this, prove (9).

17. Multinomial distribution. Suppose a trial can result in precisely one of k mutually exclusive events $A_1, \cdots, A_k$ with probabilities $p_1, \cdots, p_k$, respectively, where $p_1 + \cdots + p_k = 1$. Suppose that n independent trials are performed. Show that the probability of getting x_1 A_1's, $\cdots$, x_k A_k's is

$$f(x_1, \cdots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

where $0 \leqq x_j \leqq n$, $j = 1, \cdots, k$, and $x_1 + \cdots + x_k = n$. The distribution having this probability function is called the *multinomial distribution.*

18. A process of manufacturing screws is checked every hour by inspecting n screws selected at random from that hour's production. If one or more screws are defective, the process is halted and carefully examined. How large should n be if the manufacturer wants the probability to be about 95% that the process will be halted when 10% of the screws being produced are defective? (Assume independence of the quality of any screw from that of the other screws.)

24.8 Normal Distribution

Turning from discrete to continuous distributions, in this section we discuss the normal distribution. This is the most important continuous distribution because in applications many random variables are **normal random variables** (that is, they have a normal distribution) or they are approximately normal or can be transformed into normal random variables in a relatively simple fashion. Furthermore, the normal distribution is a useful approximation of more complicated distributions, and it also occurs in the proofs of various statistical tests.

The **normal distribution** or *Gauss distribution* is defined as the distribution with the density

$$\textbf{(1)} \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \qquad (\sigma > 0)$$

where exp is the exponential function with base $e = 2.718 \cdots$. This is simpler than it may at first look. $f(x)$ has these features (see also Fig. 519).

1. μ is the mean and σ the standard deviation.

2. $1/(\sigma\sqrt{2\pi})$ is a constant factor that makes the area under the curve of $f(x)$ from $-\infty$ to ∞ equal to 1, as it must be by (10), Sec. 24.5.

3. The curve of $f(x)$ is symmetric with respect to $x = \mu$ because the exponent is quadratic. Hence for $\mu = 0$ it is symmetric with respect to the y-axis $x = 0$ (Fig. 519, *"bell-shaped curves"*).

4. The exponential function in (1) goes to zero very fast—the faster the smaller the standard deviation σ is, as it should be (Fig. 519).

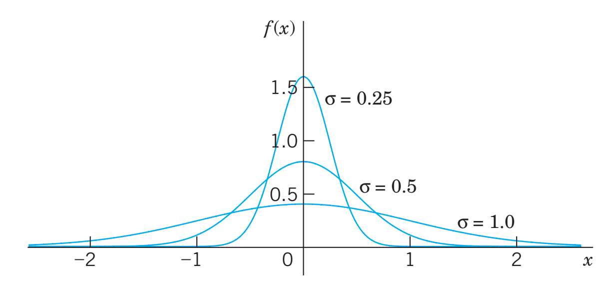


Fig. 519. Density (1) of the normal distribution with $\mu = 0$ for various values of σ

Distribution Function *F*(x)

From (7) in Sec. 24.5 and (1) we see that the normal distribution has the **distribution function**

(2)
$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left[-\frac{1}{2}\left(\frac{v-\mu}{\sigma}\right)^2\right] dv.$$

Here we needed x as the upper limit of integration and wrote v (instead of x) in the integrand.

For the corresponding **standardized normal distribution** with mean 0 and standard deviation 1 we denote $F(x)$ by $\Phi(z)$. Then we simply have from (2)

(3)
$$\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2} du.$$

This integral cannot be integrated by one of the methods of calculus. But this is no serious handicap because its values can be obtained from Table A7 in App. 5 or from your CAS. These values are needed in working with the normal distribution. The curve of $\Phi(z)$ is S-shaped. It increases monotone (why?) from 0 to 1 and intersects the vertical axis at $\frac{1}{2}$ (why?), as shown in Fig. 520.

Relation Between $F(x)$ and $\Phi(z)$. Although your CAS will give you values of $F(x)$ in (2) with any μ and σ directly, it is important to comprehend that and why any such $F(x)$ can be expressed in terms of the tabulated standard $\Phi(z)$, as follows.

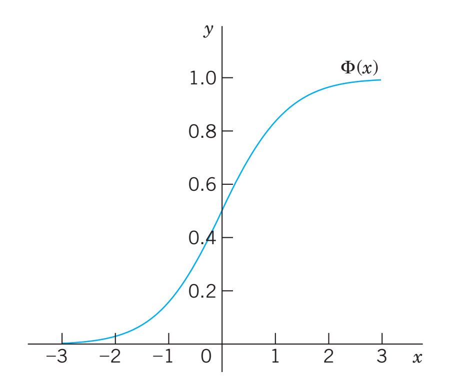


Fig. 520. Distribution function $\Phi(z)$ of the normal distribution with mean 0 and variance 1

THEOREM 1

Use of the Normal Table A7 in App. 5

The distribution function $F(x)$ of the normal distribution with any μ and σ [see (2)] *is related to the standardized distribution function $\Phi(z)$ in* (3) *by the formula*

(4)
$$F(x) = \Phi\left(\frac{x - \mu}{\sigma}\right).$$

PROOF Comparing (2) and (3) we see that we should set

$$u = \frac{v - \mu}{\sigma}. \qquad \text{Then } v = x \text{ gives} \qquad u = \frac{x - \mu}{\sigma}$$

as the new upper limit of integration. Also $v - \mu = \sigma u$, thus $dv = \sigma\, du$. Together, since σ drops out,

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{(x-\mu)/\sigma} e^{-u^2/2} \sigma\, du = \Phi\left(\frac{x - \mu}{\sigma}\right).$$ ∎

Probabilities corresponding to intervals will be needed quite frequently in statistics in Chap. 25. These are obtained as follows.

THEOREM 2

Normal Probabilities for Intervals

The probability that a normal random variable X with mean μ and standard deviation σ assume any value in an interval $a < x \leqq b$ is

(5)
$$P(a < X \leqq b) = F(b) - F(a) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

PROOF Formula (2) in Sec. 24.5 gives the first equality in (5), and (4) in this section gives the second equality. ∎

Numeric Values

In practical work with the normal distribution it is good to remember that about $\frac{2}{3}$ of all values of X to be observed will lie between $\mu \pm \sigma$, about 95% between $\mu \pm 2\sigma$, and practically all between the **three-sigma limits** $\mu \pm 3\sigma$. More precisely, by Table A7 in App. 5,

(6)

(a) $$P(\mu - \sigma < X \leqq \mu + \sigma) \approx 68\%$$

(b) $$P(\mu - 2\sigma < X \leqq \mu + 2\sigma) \approx 95.5\%$$

(c) $$P(\mu - 3\sigma < X \leqq \mu + 3\sigma) \approx 99.7\%.$$

Formulas (6a) and (6b) are illustrated in Fig. 521.

The formulas in (6) show that a value deviating from μ by more than σ, 2σ, or 3σ will occur in one of about 3, 20, and 300 trials, respectively.

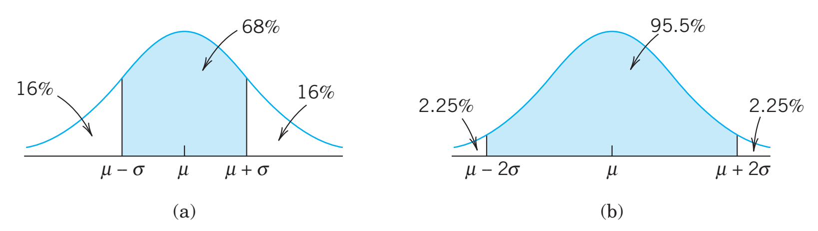


Fig. 521. Illustration of formula (6)

In tests (Chap. 25) we shall ask, conversely, for the intervals that correspond to certain given probabilities; practically most important are the probabilities of 95%, 99%, and 99.9%. For these, Table A8 in App. 5 gives the answers $\mu \pm 2\sigma$, $\mu \pm 2.6\sigma$, and $\mu \pm 3.3\sigma$, respectively. More precisely,

(7)

(a) $$P(\mu - 1.96\sigma < X \leqq \mu + 1.96\sigma) = 95\%$$

(b) $$P(\mu - 2.58\sigma < X \leqq \mu + 2.58\sigma) = 99\%$$

(c) $$P(\mu - 3.29\sigma < X \leqq \mu + 3.29\sigma) = 99.9\%.$$

Working with the Normal Tables A7 and A8 in App. 5

There are two normal tables in App. 5, Tables A7 and A8. If you want probabilities, use Table A7. If probabilities are given and corresponding intervals or x-values are wanted, use Table A8. The following examples are typical. Do them with care, verifying all values, and don't just regard them as dull exercises for your software. Make sketches of the density to see whether the results look reasonable.

EXAMPLE 1 Reading Entries from Table A7

If X is standardized normal (so that $\mu = 0, \sigma = 1$), then

$$P(X \leqq 2.44) = 0.9927 \approx 99\tfrac{1}{4}\,\%$$

$$P(X \leqq -1.16) = 1 - \Phi(1.16) = 1 - 0.8770 = 0.1230 = 12.3\%$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - 0.8413 = 0.1587) \text{ by (7), Sec. 24.3}$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(1.8) - \Phi(1.0) = 0.9641 - 0.8413 = 0.1228.$$ ■

EXAMPLE 2 Probabilities for Given Intervals, Table A7

Let X be normal with mean 0.8 and variance 4 (so that $\sigma = 2$). Then by (4) and (5)

$$P(X \leqq 2.44) = F(2.44) = \Phi\left(\frac{2.44 - 0.80}{2}\right) = \Phi(0.82) = 0.7939 \approx 80\%$$

or, if you like it better, (similarly in the other cases)

$$P(X \leqq 2.44) = P\left(\frac{X - 0.80}{2} \leqq \frac{2.44 - 0.80}{2}\right) = P(Z \leqq 0.82) = 0.7939$$

$$P(X \geqq 1) = 1 - P(X \leqq 1) = 1 - \Phi\left(\frac{1 - 0.8}{2}\right) = 1 - 0.5398 = 0.4602$$

$$P(1.0 \leqq X \leqq 1.8) = \Phi(0.5) - \Phi(0.1) = 0.6915 - 0.5398 = 0.1517.$$ ■

EXAMPLE 3 Unknown Values *c* for Given Probabilities, Table A8

Let X be normal with mean 5 and variance 0.04 (hence standard deviation 0.2). Find c or k corresponding to the given probability

$$P(X \leqq c) = 95\%, \qquad \Phi\left(\frac{c-5}{0.2}\right) = 95\%, \qquad \frac{c-5}{0.2} = 1.645, \qquad c = 5.329$$

$$P(5 - k \leqq X \leqq 5 + k) = 90\%, \qquad 5 + k = 5.329 \qquad \text{(as before; why?)}$$

$$P(X \geqq c) = 1\%, \qquad \text{thus } P(X \leqq c) = 99\%, \qquad \frac{c-5}{0.2} = 2.326, \qquad c = 5.465.$$

■

EXAMPLE 4 Defectives

In a production of iron rods let the diameter X be normally distributed with mean 2 in. and standard deviation 0.008 in.

(a) What percentage of defectives can we expect if we set the tolerance limits at 2 ± 0.02 in.?

(b) How should we set the tolerance limits to allow for 4% defectives?

Solution. **(a)** $1\frac{1}{4}\%$ because from (5) and Table A7 we obtain for the complementary event the probability

$$\begin{aligned} P(1.98 \leqq X \leqq 2.02) &= \Phi\left(\frac{2.02-2.00}{0.008}\right) - \Phi\left(\frac{1.98-2.00}{0.008}\right) \\ &= \Phi(2.5) - \Phi(-2.5) \\ &= 0.9938 - (1 - 0.9938) \\ &= 0.9876 \\ &= 98\tfrac{3}{4}\%. \end{aligned}$$

(b) 2 ± 0.0164 because, for the complementary event, we have

$$0.96 = P(2 - c \leqq X \leqq 2 + c)$$

or

$$0.98 = P(X \leqq 2 + c)$$

so that Table A8 gives

$$0.98 = \Phi\left(\frac{2+c-2}{0.008}\right),$$

$$\frac{2+c-2}{0.008} = 2.054, \qquad c = 0.0164.$$

■

Normal Approximation of the Binomial Distribution

The probability function of the binomial distribution is (Sec. 24.7)

$$f(x) = \binom{n}{x} p^x q^{n-x} \qquad (x = 0, 1, \cdots, n). \tag{8}$$

If n is large, the binomial coefficients and powers become very inconvenient. It is of great practical (and theoretical) importance that, in this case, the normal distribution provides a good approximation of the binomial distribution, according to the following theorem, one of the most important theorems in all probability theory.

THEOREM 3

Limit Theorem of De Moivre and Laplace

For large n,

$$f(x) \sim f^*(x) \qquad (x = 0, 1, \cdots, n). \tag{9}$$

Here f is given by (8). *The function*

$$f^*(x) = \frac{1}{\sqrt{2\pi}\sqrt{npq}} e^{-z^2/2}, \qquad z = \frac{x - np}{\sqrt{npq}} \tag{10}$$

is the density of the normal distribution with mean $\mu = np$ *and variance* $\sigma^2 = npq$ (*the mean and variance of the binomial distribution*). *The symbol* $\sim$ (read **asymptotically equal**) *means that the ratio of both sides approaches* 1 *as n approaches* ∞. *Furthermore, for any nonnegative integers a and b* ($> a$),

$$P(a \leqq X \leqq b) = \sum_{x=a}^{b} \binom{n}{x} p^x q^{n-x} \sim \Phi(\beta) - \Phi(\alpha),$$
$$\alpha = \frac{a - np - 0.5}{\sqrt{npq}}, \qquad \beta = \frac{b - np + 0.5}{\sqrt{npq}}. \tag{11}$$

A proof of this theorem can be found in [G3] listed in App. 1. The proof shows that the term 0.5 in α and β is a correction caused by the change from a discrete to a continuous distribution.

PROBLEM SET 24.8

1. Let X be normal with mean 10 and variance 4. Find $P(X > 12)$, $P(X < 10)$, $P(X < 11)$, $P(9 < X < 13)$.
2. Let X be normal with mean 105 and variance 25. Find $P(X \leqq 112.5)$, $P(x > 100)$, $P(110.5 < X < 111.25)$.
3. Let X be normal with mean 50 and variance 9. Determine c such that $P(X < c) = 5\%$, $P(X > c) = 1\%$, $P(50 - c < X < 50 + c) = 50\%$.
4. Let X be normal with mean 3.6 and variance 0.01. Find c such that $P(X \leqq c) = 50\%$, $P(X > c) = 10\%$, $P(-c < X - 3.6 \leqq c) = 99.9\%$.
5. If the lifetime X of a certain kind of automobile battery is normally distributed with a mean of 5 years and a standard deviation of 1 year, and the manufacturer wishes to guarantee the battery for 4 years, what percentage of the batteries will he have to replace under the guarantee?
6. If the standard deviation in Prob. 5 were smaller, would that percentage be larger or smaller?
7. A manufacturer knows from experience that the resistance of resistors he produces is normal with mean $\mu = 150\ \Omega$ and standard deviation $\sigma = 5\ \Omega$. What percentage of the resistors will have resistance between 148 Ω and 152 Ω? Between 140 Ω and 160 Ω?
8. The breaking strength X [kg] of a certain type of plastic block is normally distributed with a mean of 1500 kg and a standard deviation of 50 kg. What is the maximum load such that we can expect no more than 5% of the blocks to break?
9. If the mathematics scores of the SAT college entrance exams are normal with mean 480 and standard deviation 100 (these are about the actual values over the past years) and if some college sets 500 as the minimum score for new students, what percent of students would not reach that score?
10. A producer sells electric bulbs in cartons of 1000 bulbs. Using (11), find the probability that any given carton contains not more than 1% defective bulbs, assuming the production process to be a Bernoulli experiment with $p = 1\%$ (= probability that any given bulb will be defective). First guess. Then calculate.

11. If sick-leave time X used by employees of a company in one month is (very roughly) normal with mean 1000 hours and standard deviation 100 hours, how much time t should be budgeted for sick leave during the next month if t is to be exceeded with probability of only 20%?

12. If the monthly machine repair and maintenance cost X in a certain factory is known to be normal with mean \$12,000 and standard deviation \$2000, what is the probability that the repair cost for the next month will exceed the budgeted amount of \$15,000?

13. If the resistance X of certain wires in an electrical network is normal with mean $0.01\ \Omega$ and standard deviation $0.001\ \Omega$, how many of 1000 wires will meet the specification that they have resistance between 0.009 and $0.011\ \Omega$?

14. TEAM PROJECT. Normal Distribution. (a) Derive the formulas in (6) and (7) from the appropriate normal table.

(b) Show that $\Phi(-z) = 1 - \Phi(z)$. Give an example.

(c) Find the points of inflection of the curve of (1).

(d) Considering $\Phi^2(\infty)$ and introducing polar coordinates in the double integral (a standard trick worth remembering), prove

$$(12) \qquad \Phi(\infty) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u^2/2}\, du = 1.$$

(e) Show that σ in (1) is indeed the standard deviation of the normal distribution. [Use (12).]

(f) Bernoulli's law of large numbers. In an experiment let an event A have probability p $(0 < p < 1)$, and let X be the number of times A happens in n independent trials. Show that for any given $\epsilon > 0$,

$$P\left(\left|\frac{X}{n} - p\right| \leqq \epsilon\right) \rightarrow 1 \qquad \text{as } n \rightarrow \infty.$$

(g) Transformation. If X is normal with mean μ and variance σ^2, show that $X^* = c_1 X + c_2$ $(c_1 > 0)$ is normal with mean $\mu^* = c_1\mu + c_2$ and variance $\sigma^{*2} = c_1^2\sigma^2$.

15. WRITING PROJECT. Use of Tables, Use of CAS. Give a systematic discussion of the use of Tables A7 and A8 for obtaining $P(X < b)$, $P(X > a)$, $P(a < X < b)$, $P(X < c) = k$, $P(X > c) = k$, as well as $P(\mu - c < X < \mu + c) = k$; include simple examples. If you have a CAS, describe to what extent it makes the use of those tables superfluous; give examples.

24.9 Distributions of Several Random Variables

Distributions of two or more random variables are of interest for two reasons:

1. They occur in experiments in which we observe several random variables, for example, carbon content X and hardness Y of steel, amount of fertilizer X and yield of corn Y, height X_1, weight X_2, and blood pressure X_3 of persons, and so on.

2. They will be needed in the mathematical justification of the methods of statistics in Chap. 25.

In this section we consider two random variables X and Y or, as we also say, a **two-dimensional random variable** (X, Y). For (X, Y) the outcome of a trial is a pair of numbers $X = x$, $Y = y$, briefly $(X, Y) = (x, y)$, which we can plot as a point in the XY-plane.

The **two-dimensional probability distribution** of the random variable (X, Y) is given by the **distribution function**

$$\textbf{(1)} \qquad F(x, y) = P(X \leqq x, Y \leqq y).$$

This is the probability that in a trial, X will assume any value not greater than x and in the same trial, Y will assume any value not greater than y. This corresponds to the blue region in Fig. 522, which extends to $-\infty$ to the left and below. $F(x, y)$ determines the

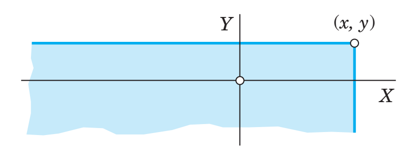


Fig. 522. Formula (1)

probability distribution uniquely, because in analogy to formula (2) in Sec. 24.5, that is, $P(a < X \leqq b) = F(b) - F(a)$, we now have for a rectangle (see Prob. 16)

$$(2) \quad P(a_1 < X \leqq b_1, \quad a_2 < Y \leqq b_2) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2).$$

As before, in the two-dimensional case we shall also have discrete and continuous random variables and distributions.

Discrete Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5), we call (X, Y) and its distribution **discrete** if (X, Y) can assume only finitely many or at most countably infinitely many pairs of values $(x_1, y_1), (x_2, y_2), \cdots$ with positive probabilities, whereas the probability for any domain containing none of those values of (X, Y) is zero.

Let (x_i, y_j) be any of those pairs and let $P(X = x_i, Y = y_j) = p_{ij}$ (where we admit that p_{ij} may be 0 for certain pairs of subscripts i, j). Then we define the **probability function** $f(x, y)$ of (X, Y) by

$$(3) \qquad f(x, y) = p_{ij} \quad \text{if} \quad x = x_i, y = y_j \qquad \text{and} \qquad f(x, y) = 0 \quad \text{otherwise;}$$

here, $i = 1, 2, \cdots$ and $j = 1, 2, \cdots$ independently. In analogy to (4), Sec. 24.5, we now have for the distribution function the formula

$$\textbf{(4)} \qquad F(x, y) = \sum_{x_i \leqq x} \sum_{y_j \leqq y} f(x_i, y_j).$$

Instead of (6) in Sec. 24.5 we now have the condition

$$(5) \qquad \sum_i \sum_j f(x_i, y_j) = 1.$$

EXAMPLE 1 Two-Dimensional Discrete Distribution

If we simultaneously toss a dime and a nickel and consider

X = *Number of heads the dime turns up,*

Y = *Number of heads the nickel turns up,*

then X and Y can have the values 0 or 1, and the probability function is

$$f(0, 0) = f(1, 0) = f(0, 1) = f(1, 1) = \tfrac{1}{4}, \quad f(x, y) = 0 \text{ otherwise.}$$ ■

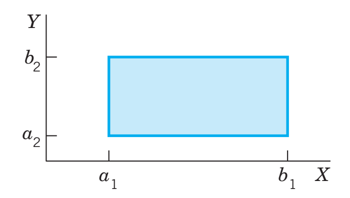


Fig. 523. Notion of a two-dimensional distribution

Continuous Two-Dimensional Distributions

In analogy to the case of a single random variable (Sec. 24.5) we call (X, Y) and its distribution **continuous** if the corresponding distribution function $F(x, y)$ can be given by a double integral

$$\textbf{(6)} \qquad F(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(x^*, y^*)\, dx^*\, dy^*$$

whose integrand f, called the **density** of (X, Y), is nonnegative everywhere, and is continuous, possibly except on finitely many curves.

From (6) we obtain the probability that (X, Y) assume any value in a rectangle (Fig. 523) given by the formula

$$(7) \qquad P(a_1 < X \leqq b_1,\quad a_2 < Y \leqq b_2) = \int_{a_2}^{b_2} \int_{a_1}^{b_1} f(x, y)\, dx\, dy.$$

EXAMPLE 2 Two-Dimensional Uniform Distribution in a Rectangle

Let R be the rectangle $\alpha_1 < x \leqq \beta_1$, $\alpha_2 < y \leqq \beta_2$. The density (see Fig. 524)

$$(8) \qquad f(x, y) = 1/k \quad \text{if } (x, y) \text{ is in } R, \qquad f(x, y) = 0 \text{ otherwise}$$

defines the so-called **uniform distribution** *in the rectangle R;* here $k = (\beta_1 - \alpha_1)(\beta_2 - \alpha_2)$ is the area of R. The distribution function is shown in Fig. 525. ■

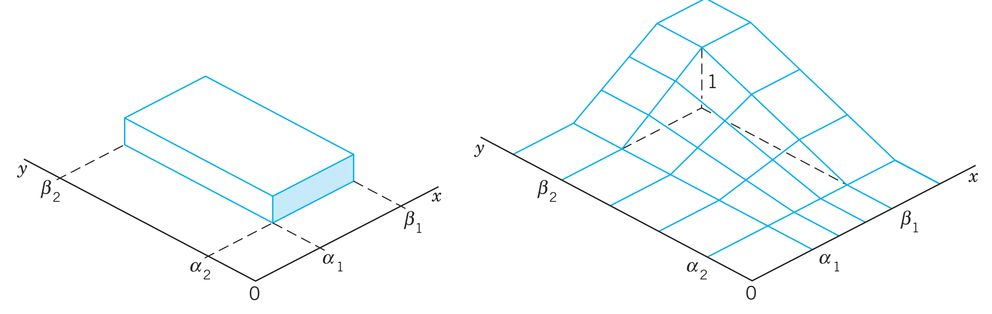


Fig. 524. Density function (8) of the uniform distribution

Fig. 525. Distribution function of the uniform distribution defined by (8)

Marginal Distributions of a Discrete Distribution

This is a rather natural idea, without counterpart for a single random variable. It amounts to being interested only in one of the two variables in (X, Y), say, X, and asking for its distribution, called the **marginal distribution** of X in (X, Y). So we ask for the probability

$P(X = x, Y \text{ arbitrary})$. Since (X, Y) is discrete, so is X. We get its probability function, call it $f_1(x)$, from the probability function $f(x, y)$ of (X, Y) by summing over y:

$$f_1(x) = P(X = x, Y \text{ arbitrary}) = \sum_{y} f(x, y) \tag{9}$$

where we sum all the values of $f(x, y)$ that are not 0 for that x.

From (9) we see that the distribution function of the marginal distribution of X is

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) = \sum_{x^* \leqq x} f_1(x^*). \tag{10}$$

Similarly, the probability function

$$f_2(y) = P(X \text{ arbitrary}, Y \leqq y) = \sum_{x} f(x, y) \tag{11}$$

determines the **marginal distribution** of Y in (X, Y). Here we sum all the values of $f(x, y)$ that are not zero for the corresponding y. The distribution function of this marginal distribution is

$$F_2(y) = P(X \text{ arbitrary}, Y \leqq y) = \sum_{y^* \leqq y} f_2(y^*). \tag{12}$$

EXAMPLE 3 Marginal Distributions of a Discrete Two-Dimensional Random Variable

In drawing 3 cards with replacement from a bridge deck let us consider

$$(X, Y), \qquad X = \textit{Number of queens}, \qquad Y = \textit{Number of kings or aces}.$$

The deck has 52 cards. These include 4 queens, 4 kings, and 4 aces. Hence in a single trial a queen has probability $\frac{4}{52} = \frac{1}{13}$ and a king or ace $\frac{8}{52} = \frac{2}{13}$. This gives the probability function of (X, Y),

$$f(x, y) = \frac{3!}{x!y!(3 - x - y)!}\left(\frac{1}{13}\right)^x\left(\frac{2}{13}\right)^y\left(\frac{10}{13}\right)^{3-x-y} \qquad (x + y \leqq 3)$$

and $f(x, y) = 0$ otherwise. Table 24.1 shows in the center the values of $f(x, y)$ and on the right and lower margins the values of the probability functions $f_1(x)$ and $f_2(y)$ of the marginal distributions of X and Y, respectively. ■

Table 24.1 Values of the Probability Functions $f(x, y)$, $f_1(x)$, $f_2(y)$ in Drawing Three Cards with Replacement from a Bridge Deck, where X is the Number of Queens Drawn and Y is the Number of Kings or Aces Drawn

x \ y	0	1	2	3	$f_1(x)$
0	$\frac{1000}{2197}$	$\frac{600}{2197}$	$\frac{120}{2197}$	$\frac{8}{2197}$	$\frac{1728}{2197}$
1	$\frac{300}{2197}$	$\frac{120}{2197}$	$\frac{12}{2197}$	0	$\frac{432}{2197}$
2	$\frac{30}{2197}$	$\frac{6}{2197}$	0	0	$\frac{36}{2197}$
3	$\frac{1}{2197}$	0	0	0	$\frac{1}{2197}$
$f_2(y)$	$\frac{1331}{2197}$	$\frac{726}{2197}$	$\frac{132}{2197}$	$\frac{8}{2197}$	

Marginal Distributions of a Continuous Distribution

This is conceptually the same as for discrete distributions, with probability functions and sums replaced by densities and integrals. For a continuous random variable (X, Y) with density $f(x, y)$ we now have the **marginal distribution** of X in (X, Y), defined by the distribution function

$$F_1(x) = P(X \leqq x, -\infty < Y < \infty) = \int_{-\infty}^{x} f_1(x^*)\, dx^* \tag{13}$$

with the density f_1 of X obtained from $f(x, y)$ by integration over y,

$$f_1(x) = \int_{-\infty}^{\infty} f(x, y)\, dy. \tag{14}$$

Interchanging the roles of X and Y, we obtain the **marginal distribution** of Y in (X, Y) with the distribution function

$$F_2(y) = P(-\infty < X < \infty, Y \leqq y) = \int_{-\infty}^{y} f_2(y^*)\, dy^* \tag{15}$$

and density

$$f_2(y) = \int_{-\infty}^{\infty} f(x, y)\, dx. \tag{16}$$

Independence of Random Variables

X and Y in a (discrete or continuous) random variable (X, Y) are said to be **independent** if

$$F(x, y) = F_1(x)F_2(y) \tag{17}$$

holds for all (x, y). Otherwise these random variables are said to be **dependent**. These definitions are suggested by the corresponding definitions for events in Sec. 24.3.

Necessary and sufficient for independence is

$$f(x, y) = f_1(x)f_2(y) \tag{18}$$

for all x and y. Here the f's are the above probability functions if (X, Y) is discrete or those densities if (X, Y) is continuous. (See Prob. 20.)

EXAMPLE 4 Independence and Dependence

In tossing a dime and a nickel, X = *Number of heads on the dime*, Y = *Number of heads on the nickel* may assume the values 0 or 1 and are independent. The random variables in Table 24.1 are dependent. ■

Extension of Independence to *n*-Dimensional Random Variables. This will be needed throughout Chap. 25. The distribution of such a random variable $\mathbf{X} = (X_1, \cdots, X_n)$ is determined by a **distribution function** of the form

$$F(x_1, \cdots, x_n) = P(X_1 \leqq x_1, \cdots, X_n \leqq x_n).$$

The random variables $X_1, \cdots, X_n$ are said to be **independent** if

$$F(x_1, \cdots, x_n) = F_1(x_1)F_2(x_2) \cdots F_n(x_n) \tag{19}$$

for all $(x_1, \cdots, x_n)$. Here $F_j(x_j)$ is the distribution function of the marginal distribution of X_j in $\mathbf{X}$, that is,

$$F_j(x_j) = P(X_j \leqq x_j, X_k \text{ arbitrary}, k = 1, \cdots, n, k \neq j).$$

Otherwise these random variables are said to be **dependent**.

Functions of Random Variables

When $n = 2$, we write $X_1 = X$, $X_2 = Y$, $x_1 = x$, $x_2 = y$. Taking a nonconstant continuous function $g(x, y)$ defined for all x, y, we obtain a random variable $Z = g(X, Y)$. For example, if we roll two dice and X and Y are the numbers the dice turn up in a trial, then $Z = X + Y$ is the sum of those two numbers (see Fig. 514 in Sec. 24.5).

In the case of a ***discrete*** random variable (X, Y) we may obtain the probability function $f(z)$ of $Z = g(X, Y)$ by summing all $f(x, y)$ for which $g(x, y)$ equals the value of z considered; thus

$$f(z) = P(Z = z) = \sum\sum_{g(x,y)=z} f(x, y). \tag{20}$$

Hence the distribution function of Z is

$$F(z) = P(Z \leqq z) = \sum\sum_{g(x,y)\leqq z} f(x, y) \tag{21}$$

where we sum all values of $f(x, y)$ for which $g(x, y) \leqq z$.

In the case of a ***continuous*** random variable (X, Y) we similarly have

$$F(z) = P(Z \leqq z) = \iint_{g(x,y)\leqq z} f(x, y)\, dx\, dy \tag{22}$$

where for each z we integrate the density $f(x, y)$ of (X, Y) over the region $g(x, y) \leqq z$ in the xy-plane, the boundary curve of this region being $g(x, y) = z$.

Addition of Means

The number

$$(23) \qquad E(g(X, Y)) = \begin{cases} \displaystyle\sum_x \sum_y g(x, y) f(x, y) & [(X, Y) \text{ discrete}] \\ \displaystyle\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f(x, y)\, dx\, dy & [(X, Y) \text{ continuous}] \end{cases}$$

is called the *mathematical expectation* or, briefly, the **expectation of** $g(X, Y)$. Here it is assumed that the double series converges absolutely and the integral of $|g(x, y)| f(x, y)$ over the xy-plane exists (is finite). Since summation and integration are linear processes, we have from (23)

$$(24) \qquad E(ag(X, Y) + bh(X, Y)) = aE(g(X, Y)) + bE(h(X, Y)).$$

An important special case is

$$E(X + Y) = E(X) + E(Y),$$

and by induction we have the following result.

THEOREM 1

Addition of Means

The mean (expectation) of a sum of random variables equals the sum of the means (expectations), that is,

$$\textbf{(25)} \qquad E(X_1 + X_2 + \cdots + X_n) = E(X_1) + E(X_2) + \cdots + E(X_n).$$

Furthermore, we readily obtain

THEROEM 2

Multiplication of Means

The mean (expectation) of the product of ***independent*** *random variables equals the product of the means (expectations), that is,*

$$(26) \qquad E(X_1 X_2 \cdots X_n) = E(X_1)E(X_2) \cdots E(X_n).$$

PROOF If X and Y are independent random variables (both discrete or both continuous), then $E(XY) = E(X)E(Y)$. In fact, in the discrete case we have

$$E(XY) = \sum_x \sum_y xyf(x, y) = \sum_x xf_1(x) \sum_y yf_2(y) = E(X)E(Y),$$

and in the continuous case the proof of the relation is similar. Extension to n independent random variables gives (26), and Theorem 2 is proved. ■

Addition of Variances

This is another matter of practical importance that we shall need. As before, let $Z = X + Y$ and denote the mean and variance of Z by μ and σ^2. Then we first have (see Team Project 20(a) in Problem Set 24.6)

$$\sigma^2 = E([Z - \mu]^2) = E(Z^2) - [E(Z)]^2.$$

From (24) we see that the first term on the right equals

$$E(Z^2) = E(X^2 + 2XY + Y^2) = E(X^2) + 2E(XY) + E(Y^2).$$

For the second term on the right we obtain from Theorem 1

$$[E(Z)]^2 = [E(X) + E(Y)]^2 = [E(X)]^2 + 2E(X)E(Y) + [E(Y)]^2.$$

By substituting these expressions into the formula for σ^2 we have

$$\begin{aligned}\sigma^2 = E(X^2) - [E(X)]^2 + E(Y^2) - [E(Y)]^2 \\ + 2[E(XY) - E(X)E(Y)].\end{aligned}$$

From Team Project 20, Sec. 24.6, we see that the expression in the first line on the right is the sum of the variances of X and Y, which we denote by σ_1^2 and σ_2^2, respectively. The quantity in the second line (except for the factor 2) is

$$\sigma_{XY} = E(XY) - E(X)E(Y) \qquad \textbf{(27)}$$

and is called the **covariance** of X and Y. Consequently, our result is

$$\sigma^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_{XY}. \qquad (28)$$

If X and Y are independent, then

$$E(XY) = E(X)E(Y);$$

hence $\sigma_{XY} = 0$, and

$$\sigma^2 = \sigma_1^2 + \sigma_2^2. \qquad (29)$$

Extension to more than two variables gives the basic

THEOREM 3

Addition of Variances

The variance of the sum of ***independent*** *random variables equals the sum of the variances of these variables.*

CAUTION! In the numerous applications of Theorems 1 and 3 we must always remember that Theorem 3 holds only for ***independent*** variables.

This is the end of Chap. 24 on probability theory. Most of the concepts, methods, and special distributions discussed in this chapter will play a fundamental role in the next chapter, which deals with methods of **statistical inference**, that is, conclusions from samples to populations, whose unknown properties we want to know and try to discover by looking at suitable properties of samples that we have obtained.

PROBLEM SET 24.9

1. Let $f(x, y) = k$ when $8 \leqq x \leqq 12$ and $0 \leqq y \leqq 2$ and zero elsewhere. Find k. Find $P(X \leqq 11, 1 \leqq Y \leqq 1.5)$ and $P(9 \leqq X \leqq 13, Y \leqq 1)$.

2. Find $P(X > 4, Y > 4)$ and $P(X \leqq 1, Y \leqq 1)$ if (X, Y) has the density $f(x, y) = \frac{1}{32}$ if $x \geqq 0, y \geqq 0, x + y \leqq 8$.

3. Let $f(x, y) = k$ if $x > 0, y > 0, x + y < 3$ and 0 otherwise. Find k. Sketch $f(x, y)$. Find $P(X + Y \leqq 1)$, $P(Y > X)$.

4. Find the density of the marginal distribution of X in Prob. 2.

5. Find the density of the marginal distribution of Y in Fig. 524.

6. If certain sheets of wrapping paper have a mean weight of 10 g each, with a standard deviation of 0.05 g, what are the mean weight and standard deviation of a pack of 10,000 sheets?

7. What are the mean thickness and the standard deviation of transformer cores each consisting of 50 layers of sheet metal and 49 insulating paper layers if the metal sheets have mean thickness 0.5 mm each with a standard deviation of 0.05 mm and the paper layers have mean 0.05 mm each with a standard deviation of 0.02 mm?

8. Let X [cm] and Y [cm] be the diameters of a pin and hole, respectively. Suppose that (X, Y) has the density

$$f(x, y) = 625 \quad \text{if} \quad 0.98 < x < 1.02, \quad 1.00 < y < 1.04$$

and 0 otherwise. **(a)** Find the marginal distributions. **(b)** What is the probability that a pin chosen at random will fit a hole whose diameter is 1.00?

9. Using Theorems 1 and 3, obtain the formulas for the mean and the variance of the binomial distribution.

10. Using Theorem 1, obtain the formula for the mean of the hypergeometric distribution. Can you use Theorem 3 to obtain the variance of that distribution?

11. A 5-gear assembly is put together with spacers between the gears. The mean thickness of the gears is 5.020 cm with a standard deviation of 0.003 cm. The mean thickness of the spacers is 0.040 cm with a standard deviation of 0.002 cm. Find the mean and standard deviation of the assembled units consisting of 5 randomly selected gears and 4 randomly selected spacers.

12. If the mean weight of certain (empty) containers is 5 lb the standard deviation is 0.2 lb, and if the filling of the containers has mean weight 100 lb and standard deviation 0.5 lb, what are the mean weight and the standard deviation of filled containers?

13. Find $P(X > Y)$ when (X, Y) has the density

$$f(x, y) = 0.25e^{-0.5(x+y)} \quad \text{if} \quad x \geqq 0, y \geqq 0$$

and 0 otherwise.

14. An electronic device consists of two components. Let X and Y [years] be the times to failure of the first and second components, respectively. Assume that (X, Y) has the density $f(x, y) = 4e^{-2(x+y)}$ if $x > 0$ and $y > 0$ and 0 otherwise. **(a)** Are X and Y dependent or independent? **(b)** Find the densities of the marginal distributions. **(c)** What is the probability that the first component will have a lifetime of 2 years or longer?

15. Give an example of two different discrete distributions that have the same marginal distributions.

16. Prove (2).

17. Let (X, Y) have the probability function

$$f(0, 0) = f(1, 1) = \tfrac{1}{8},$$
$$f(0, 1) = f(1, 0) = \tfrac{3}{8}.$$

Are X and Y independent?

18. Let (X, Y) have the density

$$f(x, y) = k \text{ if } x^2 + y^2 < 1$$

and 0 otherwise. Determine k. Find the densities of the marginal distributions. Find the probability

$$P(X^2 + Y^2 < \tfrac{1}{4}).$$

19. Show that the random variables with the densities

$$f(x, y) = x + y$$

and

$$g(x, y) = (x + \tfrac{1}{2})(y + \tfrac{1}{2})$$

if $0 \leqq x \leqq 1, 0 \leqq y \leqq 1$ and $f(x, y) = 0$ and $g(x, y) = 0$ elsewhere, have the same marginal distribution.

20. Prove the statement involving (18).

CHAPTER 24 REVIEW QUESTIONS AND PROBLEMS

1. What are stem-and-leaf plots? Boxplots? Histograms? Compare their advantages.

2. What properties of data are measured by the mean? The median? The standard deviation? The variance?

3. What do we mean by an experiment? An outcome? An event? Give examples.

4. What is a random variable? Its distribution function? Its probability function or density?

5. State the definition of probability from memory. Give simple examples.

6. What is sampling with and without replacement? What distributions are involved?

7. When is the Poisson distribution a good approximation of the binomial distribution? The normal distribution?

8. Explain the use of the tables of the normal distribution. If you have a CAS, how would you proceed without the tables?

9. State the main theorems on probability. Illustrate them by simple examples.

10. State the most important facts about distributions of two random variables and their marginal distributions.

11. Make a stem-and-leaf plot, histogram, and boxplot of the data 110, 113, 109, 118, 110, 115, 104, 111, 116, 113.

12. Same task as in Prob. 11. for the data 13.5, 13.2, 12.1, 13.6, 13.3.

13. Find the mean, standard deviation, and variance in Prob. 11.

14. Find the mean, standard deviation, and variance in Prob. 12.

15. Show that the mean always lies between the smallest and the largest data value.

16. What are the outcomes in the sample space of the experiment of simultaneously tossing three coins?

17. Plot a histogram of the data 8, 2, 4, 10 and guess $\bar{x}$ and s by inspecting the histogram. Then calculate $\bar{x}$, s^2, and s.

18. Using a Venn diagram, show that $A \subseteq B$ if and only if $A \cap B = A$.

19. Suppose that 3% of bolts made by a machine are defective, the defectives occurring at random during production. If the bolts are packaged 50 per box, what is the binomial approximation of the probability that a given box will contain $x = 0, 1, \cdots, 5$ defectives?

20. Of a lot of 12 items, 3 are defective. **(a)** Find the number of different samples of 3 items. Find the number of samples of 3 items containing **(b)** no defectives, **(c)** 1 defective, **(d)** 2 defectives, **(e)** 3 defectives.

21. Find the probability function of X = *Number of times of tossing a fair coin until the first head appears.*

22. If the life of ball bearings has the density $f(x) = ke^{-x}$ if $0 \leqq x \leqq 2$ and 0 otherwise, what is k? What is the probability $P(X \geqq 1)$?

23. Find the mean and variance of a discrete random variable X having the probability function $f(0) = \frac{1}{4}$, $f(1) = \frac{1}{2}$, $f(2) = \frac{1}{4}$.

24. Let X be normal with mean 14 and variance 4. Determine c such that $P(X \leqq c) = 95\%$, $P(X \leqq c) = 5\%$, $P(X \leqq c) = 99.5\%$.

25. Let X be normal with mean 80 and variance 9. Find $P(X > 83)$, $P(X < 81)$, $P(X < 80)$, and $P(78 < X < 82)$.

SUMMARY OF CHAPTER 24
Data Analysis. Probability Theory

A *random experiment,* briefly called **experiment**, is a process in which the result ("**outcome**") depends on "chance" (effects of factors unknown to us). Examples are games of chance with dice or cards, measuring the hardness of steel, observing weather conditions, or recording the number of accidents in a city. (Thus the word "experiment" is used here in a much wider sense than in common language.) The outcomes are regarded as points (elements) of a set S, called the **sample space**, whose subsets are called **events**. For events E we define a **probability** $P(E)$ by the axioms (Sec. 24.3)

$$
\begin{aligned}
&0 \leqq P(E) \leqq 1 \\
(1) \qquad &P(S) = 1 \\
&P(E_1 \cup E_2 \cup \cdots) = P(E_1) + P(E_2) + \cdots \qquad (E_j \cap E_k = \varnothing).
\end{aligned}
$$

These axioms are motivated by properties of frequency distributions of data (Sec. 24.1).

The complement E^c of E has the probability

$$P(E^c) = 1 - P(E). \tag{2}$$

The **conditional probability** of an event B under the condition that an event A happens is (Sec. 24.3)

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \qquad [P(A) > 0]. \tag{3}$$

Two events A and B are called **independent** if the probability of their simultaneous appearance in a trial equals the product of their probabilities, that is, if

$$P(A \cap B) = P(A)P(B). \tag{4}$$

With an experiment we associate a **random variable** X. This is a function defined on S whose values are real numbers; furthermore, X is such that the probability $P(X = a)$ with which X assumes any value a, and the probability $P(a < X \leqq b)$ with which X assumes any value in an interval $a < X \leqq b$ are defined (Sec. 24.5). The **probability distribution** of X is determined by the distribution function

$$F(x) = P(X \leqq x). \tag{5}$$

In applications there are two important kinds of random variables: those of the **discrete** type, which appear if we count (defective items, customers in a bank, etc.) and those of the **continuous** type, which appear if we measure (length, speed, temperature, weight, etc.).

A discrete random variable has a **probability function**

$$f(x) = P(X = x). \tag{6}$$

Its **mean** μ and **variance** σ^2 are (Sec. 24.6)

$$\mu = \sum_j x_j f(x_j) \qquad \text{and} \qquad \sigma^2 = \sum_j (x_j - \mu)^2 f(x_j) \tag{7}$$

where the x_j are the values for which X has a positive probability. Important discrete random variables and distributions are the **binomial, Poisson,** and **hypergeometric distributions** discussed in Sec. 24.7.

A continuous random variable has a **density**

$$f(x) = F'(x) \qquad [\text{see } (5)]. \tag{8}$$

Its mean and variance are (Sec. 24.6)

$$\mu = \int_{-\infty}^{\infty} xf(x)\,dx \qquad \text{and} \qquad \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)\,dx. \tag{9}$$

Very important is the **normal distribution** (Sec. 24.8), whose density is

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{10}$$

and whose distribution function is (Sec. 24.8; Tables A7, A8 in App. 5)

$$F(x) = \Phi\left(\frac{x-\mu}{\sigma}\right). \tag{11}$$

A **two-dimensional random variable** (X, Y) occurs if we simultaneously observe two quantities (for example, height X and weight Y of adults). Its distribution function is (Sec. 24.9)

$$F(x, y) = P(X \leqq x, Y \leqq y). \tag{12}$$

X and Y have the distribution functions (Sec. 24.9)

$$F_1(x) = P(X \leqq x, Y \text{ arbitrary}) \qquad \text{and} \qquad F_2(y) = P(x \text{ arbitrary}, Y \leqq y) \tag{13}$$

respectively; their distributions are called **marginal distributions**. If both X and Y are discrete, then (X, Y) has a probability function

$$f(x, y) = P(X = x, Y = y).$$

If both X and Y are continuous, then (X, Y) has a density $f(x, y)$.

CHAPTER 25

Mathematical Statistics

In probability theory we set up mathematical models of processes that are affected by "chance." In mathematical statistics or, briefly, **statistics**, we check these models against the observable reality. This is called **statistical inference**. It is done by **sampling**, that is, by drawing random samples, briefly called **samples**. These are sets of values from a much larger set of values that could be studied, called the **population**. An example is 10 diameters of screws drawn from a large lot of screws. Sampling is done in order to see whether a model of the population is accurate enough for practical purposes. If this is the case, the model can be used for predictions, decisions, and actions, for instance, in planning productions, buying equipment, investing in business projects, and so on.

Most important methods of statistical inference are **estimation of parameters** (Secs. 25.2), determination of **confidence intervals** (Sec. 25.3), and **hypothesis testing** (Sec. 25.4, 25.7, 25.8), with application to *quality control* (Sec. 25.5) and *acceptance sampling* (Sec. 25.6).

In the last section (25.9) we give an introduction to **regression** and **correlation analysis**, which concern experiments involving two variables.

Prerequisite: Chap. 24.
Sections that may be omitted in a shorter course: 25.5, 25.6, 25.8.
References, Answers to Problems, and Statistical Tables: App. 1 Part G, App. 2, App. 5.

25.1 Introduction. Random Sampling

Mathematical statistics consists of methods for designing and evaluating random experiments to obtain information about practical problems, such as exploring the relation between iron content and density of iron ore, the quality of raw material or manufactured products, the efficiency of air-conditioning systems, the performance of certain cars, the effect of advertising, the reactions of consumers to a new product, etc.

Random variables occur more frequently in engineering (and elsewhere) than one would think. For example, properties of mass-produced articles (screws, lightbulbs, etc.) always show **random variation**, due to small (uncontrollable!) differences in raw material or manufacturing processes. Thus the diameter of screws is a random variable X and we have *nondefective screws,* with diameter between given tolerance limits, and *defective screws,* with diameter outside those limits. We can ask for the distribution of X, for the percentage of defective screws to be expected, and for necessary improvements of the production process.

Samples are selected from populations—20 screws from a lot of 1000, 100 of 5000 voters, 8 beavers in a wildlife conservation project—because inspecting the entire population would be too expensive, time-consuming, impossible or even senseless (think

of destructive testing of lightbulbs or dynamite). To obtain meaningful conclusions, samples must be **random selections**. Each of the 1000 screws must have the same chance of being sampled (of being drawn when we sample), at least approximately. Only then will the sample mean $\bar{x} = (x_1 + \cdots + x_{20})/20$ (Sec. 24.1) of a sample of size $n = 20$ (or any other n) be a good approximation of the population mean μ (Sec. 24.6); and the accuracy of the approximation will generally improve with increasing n, as we shall see. Similarly for other parameters (standard deviation, variance, etc.).

Independent sample values will be obtained in experiments with an infinite sample space S (Sec. 24.2), certainly for the normal distribution. This is also true in sampling with replacement. It is approximately true in drawing *small* samples from a large finite population (for instance, 5 or 10 of 1000 items). However, if we sample without replacement from a small population, the effect of dependence of sample values may be considerable.

Random numbers help in obtaining samples that are in fact random selections. This is sometimes not easy to accomplish because there are many subtle factors that can bias sampling (by personal interviews, by poorly working machines, by the choice of nontypical observation conditions, etc.). Random numbers can be obtained from a **random number generator** in Maple, Mathematica, or other systems listed on p. 789. (The numbers are not truly random, as they would be produced in flipping coins or rolling dice, but are calculated by a tricky formula that produces numbers that do have practically all the essential features of true randomness. Because these numbers eventually repeat, they must not be used in cryptography, for example, where true randomness is required.)

EXAMPLE 1 Random Numbers from a Random Number Generator

To select a sample of size $n = 10$ from 80 given ball bearings, we number the bearings from 1 to 80. We then let the generator randomly produce 10 of the integers from 1 to 80 and include the bearings with the numbers obtained in our sample, for example.

44 55 53 03 52 61 67 78 39 54

or whatever.

Random numbers are also contained in (older) statistical tables. ■

Representing and processing data were considered in Sec. 24.1 in connection with frequency distributions. These are the empirical counterparts of probability distributions and helped motivating axioms and properties in probability theory. The new aspect in this chapter is **randomness**: the data are samples selected **randomly** from a population. Accordingly, we can immediately make the connection to Sec. 24.1, using stem-and-leaf plots, box plots, and histograms for representing samples graphically.

Also, we now call the mean $\bar{x}$ in (5), Sec. 24.1, the **sample mean**

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + x_2 + \cdots + x_n). \tag{1}$$

We call n the **sample size**, the variance s^2 in (6), Sec. 24.1, the **sample variance**

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n} (x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2], \tag{2}$$

and its positive square root s the **sample standard deviation**. $\bar{x}$, s^2, and s are called **parameters** *of a sample;* they will be needed throughout this chapter.

25.2 Point Estimation of Parameters

Beginning in this section, we shall discuss the most basic practical tasks in statistics and corresponding statistical methods to accomplish them. The first of them is point estimation of **parameters**, that is, of quantities appearing in distributions, such as p in the binomial distribution and μ and σ in the normal distribution.

A **point estimate** of a parameter is a number (point on the real line), which is computed from a given sample and serves as an approximation of the unknown exact value of the parameter of the population. An **interval estimate** is an interval (*"confidence interval"*) obtained from a sample; such estimates will be considered in the next section. Estimation of parameters is of great practical importance in many applications.

As an approximation of the mean μ of a population we may take the mean $\bar{x}$ of a corresponding sample. This gives the estimate $\hat{\mu} = \bar{x}$ for μ, that is,

(1)
$$\hat{\mu} = \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n)$$

where n is the sample size. Similarly, an estimate $\hat{\sigma}^2$ for the variance of a population is the variance s^2 of a corresponding sample, that is,

(2)
$$\hat{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2.$$

Clearly, (1) and (2) are estimates of parameters for distributions in which μ or σ^2 appear explicity as parameters, such as the normal and Poisson distributions. For the binomial distribution, $p = \mu/n$ [see (3) in Sec. 24.7]. From (1) we thus obtain for p the estimate

(3)
$$\hat{p} = \frac{\bar{x}}{n}.$$

We mention that (1) is a special case of the so-called **method of moments**. In this method the parameters to be estimated are expressed in terms of the moments of the distribution (see Sec. 24.6). In the resulting formulas, those moments of the distribution are replaced by the corresponding moments of the sample. This gives the estimates. Here the ***k*th moment of a sample** $x_1, \cdots, x_n$ is

$$m_k = \frac{1}{n}\sum_{j=1}^{n} x_j^k.$$

Maximum Likelihood Method

Another method for obtaining estimates is the so-called **maximum likelihood method** of R. A. Fisher [*Messenger Math.* **41** (1912), 155–160]. To explain it, we consider a discrete (or continuous) random variable X whose probability function (or density) $f(x)$ depends on a single parameter θ. We take a corresponding sample of n ***independent*** values $x_1, \cdots, x_n$. Then in the discrete case the probability that a sample of size n consists precisely of those n values is

(4) $$l = f(x_1)f(x_2) \cdots f(x_n).$$

In the continuous case the probability that the sample consists of values in the small intervals $x_j \leqq x \leqq x_j + \Delta x$ $(j = 1, 2, \cdots, n)$ is

(5) $$f(x_1)\Delta x\, f(x_2)\Delta x \cdots f(x_n)\Delta x = l(\Delta x)^n.$$

Since $f(x_j)$ depends on θ, the function l in (5) given by (4) depends on $x_1, \cdots, x_n$ and θ. We imagine $x_1, \cdots, x_n$ to be given and fixed. Then l is a function of θ, which is called the **likelihood function**. The basic idea of the maximum likelihood method is quite simple, as follows. We choose *that* approximation for the unknown value of θ for which l is as large as possible. If l is a differentiable function of θ, a necessary condition for l to have a maximum in an interval (not at the boundary) is

(6) $$\frac{\partial l}{\partial \theta} = 0.$$

(We write a *partial* derivative, because l depends also on $x_1, \cdots, x_n$.) A solution of (6) depending on $x_1, \cdots, x_n$ is called a **maximum likelihood estimate** for θ. We may replace (6) by

(7) $$\frac{\partial \ln l}{\partial \theta} = 0,$$

because $f(x_j) > 0$, a maximum of l is in general positive, and $\ln l$ is a monotone increasing function of l. This often simplifies calculations.

Several Parameters. If the distribution of X involves r parameters $\theta_1, \cdots, \theta_r$, then instead of (6) we have the r conditions $\partial l/\partial\theta_1 = 0, \cdots, \partial l/\partial\theta_r = 0$, and instead of (7) we have

(8) $$\frac{\partial \ln l}{\partial \theta_1} = 0, \qquad \cdots, \qquad \frac{\partial \ln l}{\partial \theta_r} = 0.$$

EXAMPLE 1 Normal Distribution

Find maximum likelihood estimates for $\theta_1 = \mu$ and $\theta_2 = \sigma$ in the case of the normal distribution.

Solution. From (1), Sec. 24.8, and (4) we obtain the likelihood function

$$l = \left(\frac{1}{\sqrt{2\pi}}\right)^n \left(\frac{1}{\sigma}\right)^n e^{-h} \qquad \text{where} \qquad h = \frac{1}{2\sigma^2} \sum_{j=1}^{n} (x_j - \mu)^2.$$

Taking logarithms, we have

$$\ln l = -n \ln \sqrt{2\pi} - n \ln \sigma - h.$$

The first equation in (8) is $\partial(\ln l)/\partial\mu = 0$, written out

$$\frac{\partial \ln l}{\partial \mu} = -\frac{\partial h}{\partial \mu} = \frac{1}{\sigma^2}\sum_{j=1}^{n}(x_j - \mu) = 0, \qquad \text{hence} \qquad \sum_{j=1}^{n} x_j - n\mu = 0.$$

The solution is the desired estimate $\hat{\mu}$ for μ: we find

$$\hat{\mu} = \frac{1}{n}\sum_{j=1}^{n} x_j = \bar{x}.$$

The second equation in (8) is $\partial(\ln l)/\partial\sigma = 0$, written out

$$\frac{\partial \ln l}{\partial \sigma} = -\frac{n}{\sigma} - \frac{\partial h}{\partial \sigma} = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{n}(x_j - \mu)^2 = 0.$$

Replacing μ by $\hat{\mu}$ and solving for σ^2, we obtain the estimate

$$\tilde{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_j - \bar{x})^2$$

which we shall use in Sec. 25.7. Note that this differs from (2). We cannot discuss criteria for the goodness of estimates but want to mention that for small n, formula (2) is preferable. ■

PROBLEM SET 25.2

1. **Normal distribution.** Apply the maximum likelihood method to the normal distribution with $\mu = 0$.
2. Find the maximum likelihood estimate for the parameter μ of a normal distribution with known variance $\sigma^2 = \sigma_0^2 = 16$.
3. **Poisson distribution.** Derive the maximum likelihood estimator for μ. Apply it to the sample (10, 25, 26, 17, 10, 4), giving numbers of minutes with 0–10, 11–20, 21–30, 31–40, 41–50, more than 50 fliers per minute, respectively, checking in at some airport check-in.
4. **Uniform distribution.** Show that, in the case of the parameters a and b of the uniform distribution (see Sec. 24.6), the maximum likelihood estimate cannot be obtained by equating the first derivative to zero. How can we obtain maximum likelihood estimates in this case, more or less by using common sense?
5. **Binomial distribution.** Derive a maximum likelihood estimate for p.
6. Extend Prob. 5 as follows. Suppose that m times n trials were made and in the first n trials A happened k_1 times, in the second n trials A happened k_2 times, $\cdots$, in the mth n trials A happened k_m times. Find a maximum likelihood estimate of p based on this information.
7. Suppose that in Prob. 6 we made 3 times 4 trials and A happened 2, 3, 2 times, respectively. Estimate p.
8. **Geometric distribution.** Let X = *Number of independent trials until an event A occurs*. Show that X has a geometric distribution, defined by the probability function $f(x) = pq^{x-1}$, $x = 1, 2, \cdots$, where p is the probability of A in a single trial and $q = 1 - p$. Find the maximum likelihood estimate of p corresponding to a sample $x_1, x_2, \cdots, x_n$ of observed values of X.
9. In Prob. 8, show that $f(1) + f(2) + \cdots = 1$ (as it should be!). Calculate independently of Prob. 8 the maximum likelihood of p in Prob. 8 corresponding to a single observed value of X.
10. In rolling a die, suppose that we get the first "*Six*" in the 7th trial and in doing it again we get it in the 6th trial. Estimate the probability p of getting a "*Six*" in rolling that die once.
11. Find the maximum likelihood estimate of θ in the density $f(x) = \theta e^{-\theta x}$ if $x \geqq 0$ and $f(x) = 0$ if $x < 0$.
12. In Prob. 11, find the mean μ, substitute it in $f(x)$, find the maximum likelihood estimate of μ, and show that it is identical with the estimate for μ which can be obtained from that for θ in Prob. 11.

13. Compute $\hat{\theta}$ in Prob. 11 from the sample 1.9, 0.4, 0.7, 0.6, 1.4. Graph the sample distribution function $\hat{F}(x)$ and the distribution function $F(x)$ of the random variable, with $\theta = \hat{\theta}$, on the same axes. Do they agree reasonably well? (We consider goodness of fit systematically in Sec. 25.7.)

14. Do the same task as in Prob. 13 if the given sample is 0.4, 0.7, 0.2, 1.1, 0.1.

15. CAS EXPERIMENT. Maximum Likelihood Estimates. (MLEs). Find experimentally how much MLEs can differ depending on the sample size. *Hint.* Generate many samples of the same size n, e.g., of the standardized normal distribution, and record $\bar{x}$ and s^2. Then increase n.

25.3 Confidence Intervals

Confidence intervals[1] for an unknown parameter θ of some distribution (e.g., $\theta = \mu$) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ that contain θ, not with certainty but with a high probability γ, which we can choose (95% and 99% are popular). Such an interval is calculated from a sample. $\gamma = 95\%$ means probability $1 - \gamma = 5\% = \frac{1}{20}$ of being wrong—one of about 20 such intervals will not contain θ. Instead of writing $\theta_1 \leqq \theta \leqq \theta_2$, we denote this more distinctly by writing

$$\text{CONF}_{\gamma}\,\{\theta_1 \leqq \theta \leqq \theta_2\}. \tag{1}$$

Such a special symbol, CONF, seems worthwhile in order to avoid the misunderstanding that θ *must* lie between θ_1 and θ_2.

γ is called the **confidence level**, and θ_1 and θ_2 are called the **lower** and **upper confidence limits**. They depend on γ. The larger we choose γ, the smaller is the error probability $1 - \gamma$, but the longer is the confidence interval. If $\gamma \to 1$, then its length goes to infinity. The choice of γ depends on the kind of application. In taking no umbrella, a 5% chance of getting wet is not tragic. In a medical decision of life or death, a 5% chance of being wrong may be too large and a 1% chance of being wrong ($\gamma = 99\%$) may be more desirable.

Confidence intervals are more valuable than point estimates (Sec. 25.2). Indeed, we can take the midpoint of (1) as an approximation of θ and half the length of (1) as an "error bound" (not in the strict sense of numerics, but except for an error whose probability we know).

θ_1 and θ_2 in (1) are calculated from a sample $x_1, \cdots, x_n$. These are n observations of a random variable X. Now comes a **standard trick**. *We regard $x_1, \cdots, x_n$ as* ***single*** *observations of n random variables $X_1, \cdots, X_n$ (with the same distribution, namely, that of X).* Then $\theta_1 = \theta_1(x_1, \cdots, x_n)$ and $\theta_2 = \theta_2(x_1, \cdots, x_n)$ in (1) are observed values of two random variables $\Theta_1 = \Theta_1(X_1, \cdots, X_n)$ and $\Theta_2 = \Theta_2(X_1, \cdots, X_n)$. The condition (1) involving γ can now be written

$$P(\Theta_1 \leqq \theta \leqq \Theta_2) = \gamma. \tag{2}$$

Let us see what all this means in concrete practical cases.

In each case in this section we shall first state the steps of obtaining a confidence interval in the form of a table, then consider a typical example, and finally justify those steps theoretically.

---

[1]JERZY NEYMAN (1894–1981), American statistician, developed the theory of confidence intervals (*Annals of Mathematical Statistics* **6** (1935), 111–116).

Confidence Interval for μ of the Normal Distribution with Known σ^2

Table 25.1 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Known Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the corresponding c:

γ	0.90	0.95	0.99	0.999
c	1.645	1.960	2.576	3.291

Step 3. Compute the mean $\bar{x}$ of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = c\sigma/\sqrt{n}$. The confidence interval for μ is

$$\text{CONF}_\gamma \{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{3}$$

EXAMPLE 1 Confidence Interval for μ of the Normal Distribution with Known σ^2

Determine a 95% confidence interval for the mean of a normal distribution with variance $\sigma^2 = 9$, using a sample of $n = 100$ values with mean $\bar{x} = 5$.

Solution. ***Step 1.*** $\gamma = 0.95$ is required. ***Step 2.*** The corresponding c equals 1.960; see Table 25.1. ***Step 3.*** $\bar{x} = 5$ is given. ***Step 4.*** We need $k = 1.960 \cdot 3/\sqrt{100} = 0.588$. Hence $\bar{x} - k = 4.412$, $\bar{x} + k = 5.588$ and the confidence interval is $\text{CONF}_{0.95}\{4.412 \leqq \mu \leqq 5.588\}$.

This is sometimes written $\mu = 5 \pm 0.588$, but we shall not use this notation, which can be misleading. With your CAS you can determine this interval more directly. Similarly for the other examples in this section. ■

Theory for Table 25.1. The method in Table 25.1 follows from the basic

THEOREM 1 Sum of Independent Normal Random Variables

*Let $X_1, \cdots, X_n$ be **independent** normal random variables each of which has mean μ and variance σ^2. Then the following holds.*

(a) *The sum $X_1 + \cdots + X_n$ is normal with mean $n\mu$ and variance $n\sigma^2$.*

(b) The following random variable $\bar{X}$ *is normal with mean μ and variance σ^2/n.*

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n) \tag{4}$$

(c) *The following random variable Z is normal with mean* 0 *and variance* 1.

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \tag{5}$$

PROOF The statements about the mean and variance in (a) follow from Theorems 1 and 3 in Sec. 24.9. From this, and Theorem 2 in Sec. 24.6, we see that $\bar{X}$ has the mean $(1/n)n\mu = \mu$ and the variance $(1/n)^2 n\sigma^2 = \sigma^2/n$. This implies that Z has the mean 0 and variance 1, by Theorem 2(b) in Sec. 24.6. The normality of $X_1 + \cdots + X_n$ is proved in Ref. [G3] listed in App. 1. This implies the normality of (4) and (5). ■

Derivation of (3) in Table 25.1. Sampling from a normal distribution gives independent sample values (see Sec. 25.1), so that Theorem 1 applies. Hence we can choose γ and then determine c such that

$$P(-c \leqq Z \leqq c) = P\left(-c \leqq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leqq c\right) = \Phi(c) - \Phi(-c) = \gamma. \tag{6}$$

For the value $\gamma = 0.95$ we obtain $z(D) = 1.960$ from Table A8 in App. 5, as used in Example 1. For $\gamma = 0.9, 0.99, 0.999$ we get the other values of c listed in Table 25.1. Finally, all we have to do is to convert the inequality in (6) into one for μ and insert observed values obtained from the sample. We multiply $-c \leqq Z \leqq c$ by -1 and then by $\sigma/\sqrt{n}$, writing $c\sigma/\sqrt{n} = k$ (as in Table 25.1),

$$P(-c \leqq Z \leqq c) = P(c \geqq -Z \geqq -c) = P\left(c \geqq \frac{\mu - \bar{X}}{\sigma/\sqrt{n}} \geqq -c\right)$$

$$= P(k \geqq \mu - \bar{X} \geqq -k) = \gamma.$$

Adding $\bar{X}$ gives $P(\bar{X} + k \geqq \mu \geqq \bar{X} - k) = \gamma$ or

$$P(\bar{X} - k \leqq \mu \leqq \bar{X} + k) = \gamma. \tag{7}$$

Inserting the observed value $\bar{x}$ of $\bar{X}$ gives (3). Here we have regarded $x_1, \cdots, x_n$ as single observations of $X_1, \cdots, X_n$ (the standard trick!), so that $x_1 + \cdots + x_n$ is an observed value of $X_1 + \cdots + X_n$ and $\bar{x}$ is an observed value of $\bar{X}$. Note further that (7) is of the form (2) with $\Theta_1 = \bar{X} - k$ and $\Theta_2 = \bar{X} + k$. ■

EXAMPLE 2 Sample Size Needed for a Confidence Interval of Prescribed Length

How large must n be in Example 1 if we want to obtain a 95% confidence interval of length $L = 0.4$?

Solution. The interval (3) has the length $L = 2k = 2c\sigma/\sqrt{n}$. Solving for n, we obtain

$$n = (2c\sigma/L)^2.$$

In the present case the answer is $n = (2 \cdot 1.960 \cdot 3/0.4)^2 \approx 870$.

Figure 526 shows how L decreases as n increases and that for $\gamma = 99\%$ the confidence interval is substantially longer than for $\gamma = 95\%$ (and the same sample size n). ■

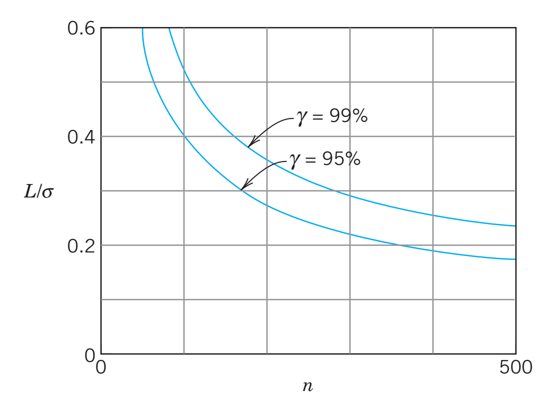


Fig. 526. Length of the confidence interval (3) (measured in multiples of σ) as a function of the sample size n for $\gamma = 95\%$ and $\gamma = 99\%$

Confidence Interval for μ of the Normal Distribution with Unknown σ^2

In practice σ^2 is frequently unknown. Then the method in Table 25.1 does not help and the whole theory changes, although the steps of determining a confidence interval for μ remain quite similar. They are shown in Table 25.2. We see that k differs from that in Table 25.1, namely, the sample standard deviation s has taken the place of the unknown standard deviation σ of the population. And c now depends on the sample size n and must be determined from Table A9 in App. 5 or from your CAS. That table lists values z for given values of the distribution function (Fig. 527)

$$F(z) = K_m \int_{-\infty}^{x} \left(1 + \frac{u^2}{m}\right)^{-(m+1)/2} du \tag{8}$$

of the ***t*-distribution**. Here, $m\ (= 1, 2, \cdots)$ is a parameter, called the **number of degrees of freedom** of the distribution (***abbreviated* d.f.**). In the present case, $m = n - 1$; see Table 25.2. The constant K_m is such that $F(\infty) = 1$. By integration it turns out that $K_m = \Gamma(\frac{1}{2}m + \frac{1}{2})/[\sqrt{m\pi}\,\Gamma(\frac{1}{2}m)]$, where Γ is the gamma function (see (24) in App. A3.1).

Table 25.2 Determination of a Confidence Interval for the Mean μ of a Normal Distribution with Unknown Variance σ^2

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{9}$$

from the table of the t-distribution with $n - 1$ degrees of freedom (Table A9 in App. 5; or use a CAS; n = sample size).

Step 3. Compute the mean $\bar{x}$ and the variance s^2 of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k = cs/\sqrt{n}$. The confidence interval is

$$\text{CONF}_\gamma\,\{\bar{x} - k \leqq \mu \leqq \bar{x} + k\}. \tag{10}$$

Figure 528 compares the curve of the density of the t-distribution with that of the normal distribution. The latter is steeper. This illustrates that Table 25.1 (which uses more information, namely, the known value of σ^2) yields shorter confidence intervals than Table 25.2. This is confirmed in Fig. 529, which also gives an idea of the gain by increasing the sample size.

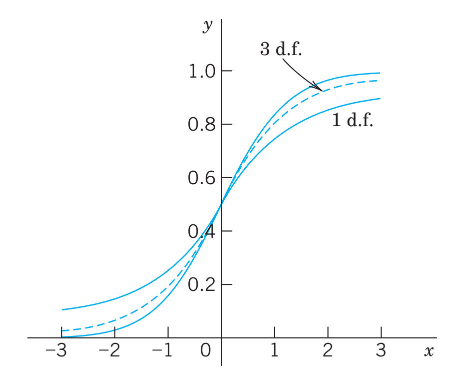


Fig. 527. Distribution functions of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution (steepest curve)

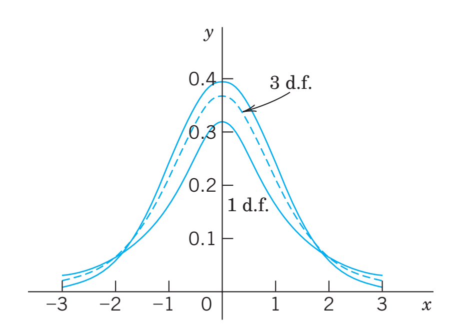


Fig. 528. Densities of the t-distribution with 1 and 3 d.f. and of the standardized normal distribution

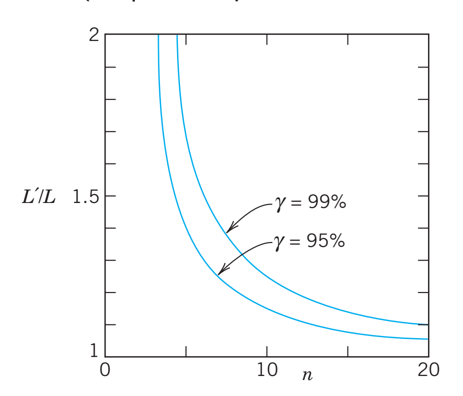


Fig. 529. Ratio of the lengths L' and L of the confidence intervals (10) and (3) with $\gamma = 95\%$ and $\gamma = 99\%$ as a function of the sample size n for equal s and σ

EXAMPLE 3 Confidence Interval for μ of the Normal Distribution with Unknown σ^2

Five independent measurements of the point of inflammation (flash point) of Diesel oil (D-2) gave the values (in °F) 144 147 146 142 144. Assuming normality, determine a 99% confidence interval for the mean.

Solution. ***Step 1.*** $\gamma = 0.99$ is required.

Step 2. $F(c) = \frac{1}{2}(1 + \gamma) = 0.995$, and Table A9 in App. 5 with $n - 1 = 4$ d.f. gives $c = 4.60$.

Step 3. $\bar{x} = 144.6, \quad s^2 = 3.8$.

Step 4. $k = \sqrt{3.8} \cdot 4.60/\sqrt{5} = 4.01$. The confidence interval is $\text{CONF}_{0.99}\,\{140.5 \leqq \mu \leqq 148.7\}$.

If the variance σ^2 were known and equal to the sample variance s^2, thus $\sigma^2 = 3.8$, then Table 25.1 would give $k = c\sigma/\sqrt{n} = 2.576\sqrt{3.8}/\sqrt{5} = 2.25$ and $\text{CONF}_{0.99}\,\{142.35 \leqq \mu \leqq 146.85\}$. We see that the present interval is almost twice as long as that obtained from Table 25.1 (with $\sigma^2 = 3.8$). Hence for small samples the difference is considerable! See also Fig. 529. ■

Theory for Table 25.2. For deriving (10) in Table 25.2 we need from Ref. [G3]

THEOREM 2

Student's *t*-Distribution

Let $X_1, \cdots, X_n$ be independent normal random variables with the same mean μ and the same variance σ^2. Then the random variable

(11)
$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

has a t-distribution [see (8)] *with $n - 1$ degrees of freedom* (d.f.); *here $\overline{X}$ is given by* (4) *and*

(12)
$$S^2 = \frac{1}{n-1} \sum_{j=1}^{n} (X_j - \overline{X})^2.$$

Derivation of (10). This is similar to the derivation of (3). We choose a number γ between 0 and 1 and determine a number c from Table A9 in App. 5 with $n - 1$ d.f. (or from a CAS) such that

(13)
$$P(-c \leqq T \leqq c) = F(c) - F(-c) = \gamma.$$

Since the t-distribution is symmetric, we have

$$F(-c) = 1 - F(c),$$

and (13) assumes the form (9). Substituting (11) into (13) and transforming the result as before, we obtain

(14)
$$P(\overline{X} - K \leqq \mu \leqq \overline{X} + K) = \gamma$$

where

$$K = cS/\sqrt{n}.$$

By inserting the observed values $\bar{x}$ of $\overline{X}$ and s^2 of S^2 into (14) we finally obtain (10). ■

Confidence Interval for the Variance σ^2 of the Normal Distribution

Table 25.3 shows the steps, which are similar to those in Tables 25.1 and 25.2.

Table 25.3 Determination of a Confidence Interval for the Variance σ^2 of a Normal Distribution, Whose Mean Need Not Be Known

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine solutions c_1 and c_2 of the equations

$$F(c_1) = \tfrac{1}{2}(1 - \gamma), \qquad F(c_2) = \tfrac{1}{2}(1 + \gamma) \tag{15}$$

from the table of the chi-square distribution with $n - 1$ degrees of freedom (Table A10 in App. 5; or use a CAS; n = sample size).

Step 3. Compute $(n - 1)s^2$, where s^2 is the variance of the sample $x_1, \cdots, x_n$.

Step 4. Compute $k_1 = (n - 1)s^2/c_1$ and $k_2 = (n - 1)s^2/c_2$. The confidence interval is

$$\text{CONF}_\gamma \,\{k_2 \leqq \sigma^2 \leqq k_1\}. \tag{16}$$

EXAMPLE 4 Confidence Interval for the Variance of the Normal Distribution

Determine a 95% confidence interval (16) for the variance, using Table 25.3 and a sample (tensile strength of sheet steel in kg/mm^2, rounded to integer values)

89 84 87 81 89 86 91 90 78 89 87 99 83 89.

Solution. ***Step 1.*** $\gamma = 0.95$ is required.

Step 2. For $n - 1 = 13$ we find

$$c_1 = 5.01 \quad \text{and} \quad c_2 = 24.74.$$

Step 3. $13s^2 = 326.9$.

Step 4. $13s^2/c_1 = 65.25$, $13s^2/c_2 = 13.21$.

The confidence interval is

$$\text{CONF}_{0.95}\,\{13.21 \leqq \sigma^2 \leqq 65.25\}.$$

This is rather large, and for obtaining a more precise result, one would need a much larger sample. ■

Theory for Table 25.3. In Table 25.1 we used the normal distribution, in Table 25.2 the t-distribution, and now we shall use the χ^2**-distribution** (*chi-square distribution*), whose distribution function is $F(z) = 0$ if $z < 0$ and

$$F(z) = C_m \int_0^z e^{-u/2} u^{(m-2)/2} \, du \qquad \text{if } z \geqq 0 \qquad \text{(Fig. 530).}$$

The parameter m (= 1, 2, $\cdots$) is called the **number of degrees of freedom** (d.f.), and

$$C_m = 1/[2^{m/2}\Gamma(\tfrac{1}{2}m)].$$

Note that the distribution is not symmetric (see also Fig. 531).

For deriving (16) in Table 25.3 we need the following theorem.

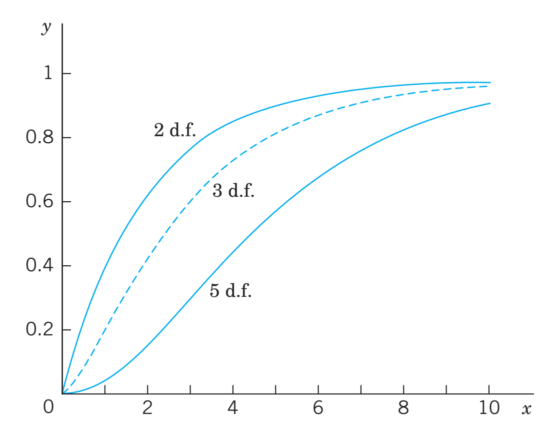


Fig. 530. Distribution function of the chi-square distribution with 2, 3, 5 d.f.

THEOREM 3

Chi-Square Distribution

Under the assumptions in Theorem 2 the random variable

$$Y = (n-1)\frac{S^2}{\sigma^2} \tag{17}$$

with S^2 given by (12) *has a chi-square distribution with $n - 1$ degrees of freedom.*

Proof in Ref. [G3], listed in App. 1.

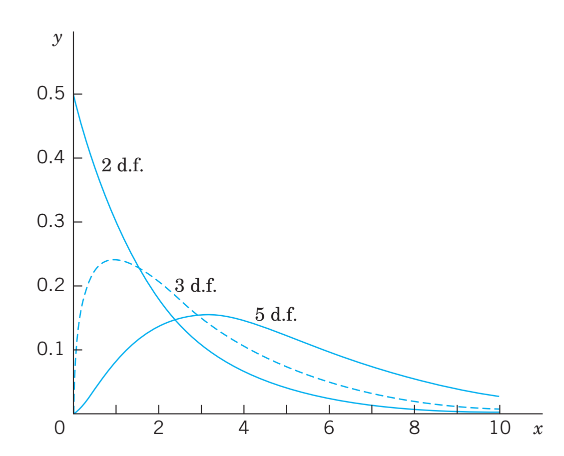


Fig. 531. Density of the chi-square distribution with 2, 3, 5 d.f.

Derivation of (16). This is similar to the derivation of (3) and (10). We choose a number γ between 0 and 1 and determine c_1 and c_2 from Table A10, App. 5, such that [see (15)]

$$P(Y \leqq c_1) = F(c_1) = \tfrac{1}{2}(1-\gamma), \qquad P(Y \leqq c_2) = F(c_2) = \tfrac{1}{2}(1+\gamma).$$

Subtraction yields

$$P(c_1 \leqq Y \leqq c_2) = P(Y \leqq c_2) - P(Y \leqq c_1) = F(c_2) - F(c_1) = \gamma.$$

Transforming $c_1 \leqq Y \leqq c_2$ with Y given by (17) into an inequality for σ^2, we obtain

$$\frac{n-1}{c_2} S^2 \leqq \sigma^2 \leqq \frac{n-1}{c_1} S^2.$$

By inserting the observed value s^2 of S^2 we obtain (16). ■

Confidence Intervals for Parameters of Other Distributions

The methods in Tables 25.1–25.3 for confidence intervals for μ and σ^2 are designed for the normal distribution. We now show that they can also be applied to other distributions if we use large samples.

We know that if $X_1, \cdots, X_n$ are independent random variables with the same mean μ and the same variance σ^2, then their sum $Y_n = X_1 + \cdots + X_n$ has the following properties.

(A) Y_n has the mean $n\mu$ and the variance $n\sigma^2$ (by Theorems 1 and 3 in Sec. 24.9).

(B) If those variables are normal, then Y_n is normal (by Theorem 1).

If those random variables are not normal, then **(B)** is not applicable. However, for large n the random variable Y_n is still *approximately* normal. This follows from the central limit theorem, which is one of the most fundamental results in probability theory.

THEOREM 4

Central Limit Theorem

Let $X_1, \cdots, X_n, \cdots$ be independent random variables that have the same distribution function and therefore the same mean μ and the same variance σ^2. Let $Y_n = X_1 + \cdots + X_n$. Then the random variable

(18)
$$Z_n = \frac{Y_n - n\mu}{\sigma\sqrt{n}}$$

is **asymptotically normal** *with mean 0 and variance 1; that is, the distribution function $F_n(x)$ of Z_n satisfies*

$$\lim_{n\to\infty} F_n(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2}\, du.$$

A proof can be found in Ref. [G3] listed in App. 1.

Hence, when applying Tables 25.1–25.3 to a nonnormal distribution, we must use ***sufficiently large samples***. As a rule of thumb, if the sample indicates that the skewness of the distribution (the asymmetry; see Team Project 20(d), Problem Set 24.6) is small, use at least $n = 20$ for the mean and at least $n = 50$ for the variance.

PROBLEM SET 25.3

1. Why are interval estimates generally more useful than point estimates?

2–6 MEAN (VARIANCE KNOWN)

2. Find a 95% confidence interval for the mean of a normal population with standard deviation 4.00 from the sample 39, 51, 49, 43, 57, 59. Does that interval get longer or shorter if we take $\gamma = 0.99$ instead of 0.95? By what factor?

3. By what factor does the length of the interval in Prob. 2 change if we double the sample size?

4. Determine a 95% confidence interval for the mean μ of a normal population with variance $\sigma^2 = 16$, using a sample of size 200 with mean 74.81.

5. What sample size would be needed for obtaining a 95% confidence interval (3) of length 2σ? Of length σ?

6. What sample size is needed to obtain a 99% confidence interval of length 2.0 for the mean of a normal population with variance 25? Use Fig. 526. Check by calculation.

MEAN (VARIANCE UNKNOWN)

7. Find a 95% confidence interval for the percentage of cars on a certain highway that have poorly adjusted brakes, using a random sample of 800 cars stopped at a roadblock on that highway, 126 of which had poorly adjusted brakes.

8. K. Pearson result. Find a 99% confidence interval for p in the binomial distribution from a classical result by K. Pearson, who in 24,000 trials of tossing a coin obtained 12,012 Heads. Do you think that the coin was fair?

9–11 Find a 99% confidence interval for the mean of a normal population from the sample:

9. Copper content (%) of brass 66, 66, 65, 64, 66, 67, 64, 65, 63, 64

10. Melting point (°C) of aluminum 660, 667, 654, 663, 662

11. Knoop hardness of diamond 9500, 9800, 9750, 9200, 9400, 9550

12. CAS EXPERIMENT. Confidence Intervals. Obtain 100 samples of size 10 of the standardized normal distribution. Calculate from them and graph the corresponding 95% confidence intervals for the mean and count how many of them do not contain 0. Does the result support the theory? Repeat the whole experiment, compare and comment.

13–17 VARIANCE

Find a 95% confidence interval for the variance of a normal population from the sample:

13. Length of 20 bolts with sample mean 20.2 cm and sample variance 0.04 cm^2

14. Carbon monoxide emission (grams per mile) of a certain type of passenger car (cruising at 55 mph): 17.3, 17.8, 18.0, 17.7, 18.2, 17.4, 17.6, 18.1

15. Mean energy (keV) of delayed neutron group (Group 3, half-life 6.2 s) for uranium U^{235} fission: a sample of 100 values with mean 442.5 and variance 9.3

16. Ultimate tensile strength (k psi) of alloy steel (Maraging H) at room temperature: 251, 255, 258, 253, 253, 252, 250, 252, 255, 256

17. The sample in Prob. 9

18. If X_1 and X_2 are independent normal random variables with mean 14 and 8 and variance 2 and 5, respectively, what distribution does $3X_1 - X_2$ have? *Hint.* Use Team Project 14(g) in Sec. 24.8.

19. A machine fills boxes weighing Y lb with X lb of salt, where X and Y are normal with mean 100 lb and 5 lb and standard deviation 1 lb and 0.5 lb, respectively. What percent of filled boxes weighing between 104 lb and 106 lb are to be expected?

20. If the weight X of bags of cement is normally distributed with a mean of 40 kg and a standard deviation of 2 kg, how many bags can a delivery truck carry so that the probability of the total load exceeding 2000 kg will be 5%?

25.4 Testing of Hypotheses. Decisions

The ideas of confidence intervals and of tests[2] are the two most important ideas in modern statistics. In a statistical **test** we make inference from sample to population through testing a **hypothesis**, resulting from experience or observations, from a theory or a quality requirement, and so on. In many cases the result of a test is used as a basis for a **decision**, for instance, to

[2]Beginning around 1930, a systematic theory of tests was developed by NEYMAN (see Sec. 25.3) and EGON SHARPE PEARSON (1895–1980), English statistician, the son of Karl Pearson (see the footnote on p. 1086).

buy (or not to buy) a certain model of car, depending on a test of the fuel efficiency (miles/gal) (and other tests, of course), to apply some medication, depending on a test of its effect; to proceed with a marketing strategy, depending on a test of consumer reactions, etc.

Let us explain such a test in terms of a typical example and introduce the corresponding standard notions of statistical testing.

EXAMPLE 1 **Test of a Hypothesis. Alternative. Significance Level α**

We want to buy 100 coils of a certain kind of wire, provided we can verify the manufacturer's claim that the wire has a breaking limit $\mu = \mu_0 = 200$ lb (or more). This is a test of the **hypothesis** (also called *null hypothesis*) $\mu = \mu_0 = 200$. We shall not buy the wire if the (statistical) test shows that actually $\mu = \mu_1 < \mu_0$, the wire is weaker, the claim does not hold. μ_1 is called the **alternative** (or *alternative hypothesis*) of the test. We shall **accept** the hypothesis if the test suggests that it is true, except for a small error probability α, called the **significance level** of the test. Otherwise we **reject** the hypothesis. Hence α is the probability of rejecting a hypothesis although it is true. The choice of α is up to us. 5% and 1% are popular values.

For the test we need a sample. We randomly select 25 coils of the wire, cut a piece from each coil, and determine the breaking limit experimentally. Suppose that this sample of $n = 25$ values of the breaking limit has the mean $\bar{x} = 197$ lb (somewhat less than the claim!) and the standard deviation $s = 6$ lb.

At this point we could only speculate whether this difference $197 - 200 = -3$ is due to randomness, is a chance effect, or whether it is **significant**, due to the actually inferior quality of the wire. To continue beyond speculation requires probability theory, as follows.

We assume that the breaking limit is normally distributed. (This assumption could be tested by the method in Sec. 25.7. Or we could remember the central limit theorem (Sec. 25.3) and take a still larger sample.) Then

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

in (11), Sec. 25.3, with $\mu = \mu_0$ has a ***t*-distribution** with $n - 1$ degrees of freedom ($n - 1 = 24$ for our sample). Also $\bar{x} = 197$ and $s = 6$ are observed values of $\bar{X}$ and S to be used later. We can now choose a significance level, say, $\alpha = 5\%$. From Table A9 in App. 5 or from a CAS we then obtain a critical value c such that $P(T \leqq c) = \alpha = 5\%$. For $P(T \leqq \tilde{c}) = 1 - \alpha = 95\%$ the table gives $\tilde{c} = 1.71$, so that $c = -\tilde{c} = -1.71$ because of the symmetry of the distribution (Fig. 532).

We now reason as follows—this is the ***crucial idea*** of the test. If the hypothesis is true, we have a chance of only α $(= 5\%)$ that we observe a value t of T (calculated from a sample) that will fall between $-\infty$ and -1.71. Hence, if we nevertheless do observe such a t, we assert that the hypothesis cannot be true and we reject it. Then we accept the alternative. If, however, $t \geqq c$, we accept the hypothesis.

A simple calculation finally gives $t = (197 - 200)/(6/\sqrt{25}) = -2.5$ as an observed value of T. Since $-2.5 < -1.71$, we reject the hypothesis (the manufacturer's claim) and accept the alternative $\mu = \mu_1 < 200$, the wire seems to be weaker than claimed. ■

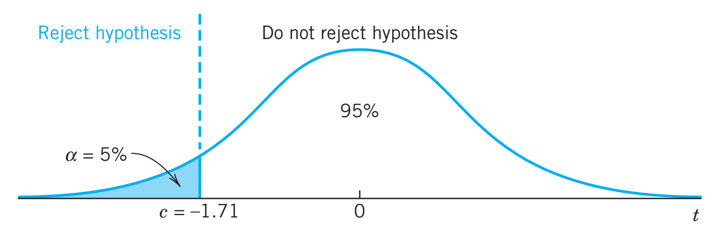


Fig. 532. *t*-distribution in Example 1

This example illustrates the ***steps of a test:***

1. Formulate the **hypothesis** $\theta = \theta_0$ to be tested. ($\theta_0 = \mu_0$ in the example.)

2. Formulate an **alternative** $\theta = \theta_1$. ($\theta_1 = \mu_1$ in the example.)

3. Choose a **significance level** α (5%, 1%, 0.1%).

4. Use a random variable $\hat{\Theta} = g(X_1, \cdots, X_n)$ whose distribution depends on the hypothesis and on the alternative, and this distribution is known in both cases. Determine

a critical value c from the distribution of $\hat{\Theta}$, assuming the hypothesis to be true. (In the example, $\hat{\Theta} = T$, and c is, obtained from $P(T \leqq c) = \alpha$.)

5. Use a sample $x_1, \cdots, x_n$ to determine an observed value $\hat{\theta} = g(x_1, \cdots, x_n)$ of $\hat{\Theta}$. (t in the example.)

6. Accept or reject the hypothesis, depending on the size of $\hat{\theta}$ relative to c. ($t < c$ in the example, rejection of the hypothesis.)

Two important facts require further discussion and careful attention. The first is the choice of an alternative. In the example, $\mu_1 < \mu_0$, but other applications may require $\mu_1 > \mu_0$ or $\mu_1 \neq \mu_0$. The second fact has to do with errors. We know that α (the significance level of the test) is the probability of *rejecting* a *true* hypothesis. And we shall discuss the probability β of *accepting* a *false* hypothesis.

One-Sided and Two-Sided Alternatives (Fig. 533)

Let θ be an unknown parameter in a distribution, and suppose that we want to test the hypothesis $\theta = \theta_0$. Then there are three main kinds of alternatives, namely,

$$\theta > \theta_0 \tag{1}$$

$$\theta < \theta_0 \tag{2}$$

$$\theta \neq \theta_0. \tag{3}$$

(1) and (2) are **one-sided alternatives**, and (3) is a **two-sided alternative**.

We call **rejection region** (or **critical region**) the region such that we reject the hypothesis if the observed value in the test falls in this region. In ① the critical c lies to the right of θ_0 because so does the alternative. Hence the rejection region extends to the right. This is called a **right-sided test**. In ② the critical c lies to the left of θ_0 (as in Example 1), the rejection region extends to the left, and we have a **left-sided test** (Fig. 533, middle part). These are **one-sided tests**. In ③ we have two rejection regions. This is called a **two-sided test** (Fig. 533, lower part).

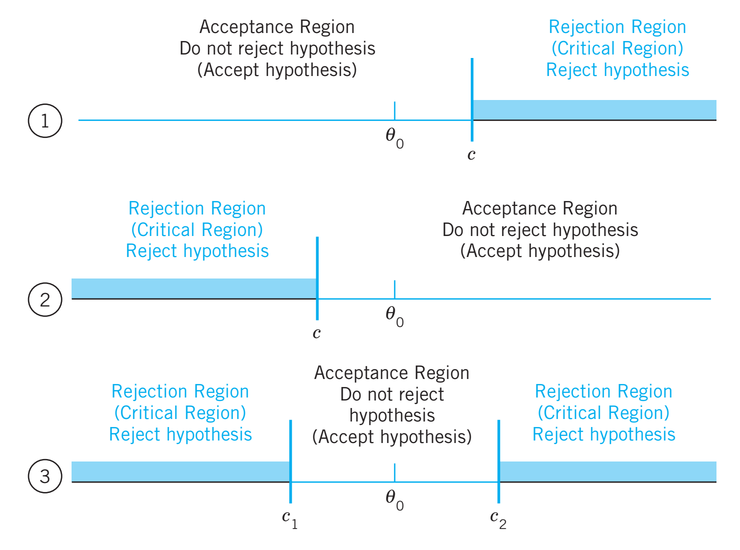


Fig. 533. Test in the case of alternative (1) (upper part of the figure), alternative (2) (middle part), and alternative (3)

All three kinds of alternatives occur in practical problems. For example, (1) may arise if θ_0 is the maximum tolerable inaccuracy of a voltmeter or some other instrument. Alternative (2) may occur in testing strength of material, as in Example 1. Finally, θ_0 in (3) may be the diameter of axle-shafts, and shafts that are too thin or too thick are equally undesirable, so that we have to watch for deviations in both directions.

Errors in Tests

Tests always involve **risks of making false decisions:**

(I) Rejecting a true hypothesis **(Type I error)**.
α = Probability of making a Type I error.

(II) Accepting a false hypothesis **(Type II error)**.
β = Probability of making a Type II error.

Clearly, we cannot avoid these errors because no absolutely certain conclusions about populations can be drawn from samples. But we show that there are ways and means of choosing suitable levels of risks, that is, of values α and β. The choice of α depends on the nature of the problem (e.g., a small risk $\alpha = 1\%$ is used if it is a matter of life or death).

Let us discuss this systematically for a test of a hypothesis $\theta = \theta_0$ against an alternative that is a single number θ_1, for simplicity. We let $\theta_1 > \theta_0$, so that we have a right-sided test. For a left-sided or a two-sided test the discussion is quite similar.

We choose a critical $c > \theta_0$ (as in the upper part of Fig. 533, by methods discussed below). From a given sample $x_1, \cdots, x_n$ we then compute a value

$$\hat{\theta} = g(x_1, \cdots, x_n)$$

with a suitable g (whose choice will be a main point of our further discussion; for instance, take $g = (x_1 + \cdots + x_n)/n$ in the case in which θ is the mean). If $\hat{\theta} > c$, we reject the hypothesis. If $\hat{\theta} \leqq c$, we accept it. Here, the value $\hat{\theta}$ can be regarded as an observed value of the random variable

$$\hat{\Theta} = g(X_1, \cdots, X_n) \tag{4}$$

because x_j may be regarded as an observed value of X_j, $j = 1, \cdots, n$. In this test there are two possibilities of making an error, as follows.

Type I Error (see Table 25.4). The hypothesis is true but is rejected (hence the alternative is accepted) because Θ assumes a value $\hat{\theta} > c$. Obviously, the probability of making such an error equals

$$P(\hat{\Theta} > c)_{\theta=\theta_0} = \alpha. \tag{5}$$

α is called the **significance level** of the test, as mentioned before.

Type II Error (see Table 25.4). The hypothesis is false but is accepted because $\hat{\Theta}$ assumes a value $\hat{\theta} \leqq c$. The probability of making such an error is denoted by β; thus

$$P(\hat{\Theta} \leqq c)_{\theta=\theta_1} = \beta. \tag{6}$$

$\eta = 1 - \beta$ is called the **power** of the test. Obviously, the power η is the probability of avoiding a Type II error.

Table 25.4 Type I and Type II Errors in Testing a Hypothesis $\theta = \theta_0$ Against an Alternative $\theta = \theta_1$

		Unknown Truth	
		$\theta = \theta_0$	$\theta = \theta_1$
Accepted	$\theta = \theta_0$	True decision $P = 1 - \alpha$	Type II error $P = \beta$
	$\theta = \theta_1$	Type 1 error $P = \alpha$	True decision $P = 1 - \beta$

Formulas (5) and (6) show that both α and β depend on c, and we would like to choose c so that these probabilities of making errors are as small as possible. But the important Figure 534 shows that these are conflicting requirements because to let α decrease we must shift c to the right, but then β increases. In practice we first choose α (5%, sometimes 1%), then determine c, and finally compute β. If β is large so that the power $\eta = 1 - \beta$ is small, we should repeat the test, choosing a larger sample, for reasons that will appear shortly.

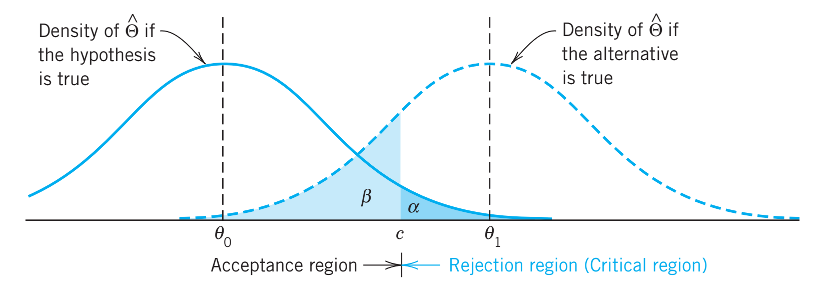


Fig. 534. Illustration of Type I and II errors in testing a hypothesis $\theta = \theta_0$ against an alternative $\theta = \theta_1$ ($> \theta_0$, right-sided test)

If the alternative is not a single number but is of the form (1)–(3), then β becomes a function of θ. This function $\beta(\theta)$ is called the **operating characteristic** (OC) of the test and its curve the **OC curve**. Clearly, in this case $\eta = 1 - \beta$ also depends on θ. This function $\eta(\theta)$ is called the **power function** of the test. (Examples will follow.)

Of course, from a test that leads to the acceptance of a certain hypothesis θ_0, it does ***not*** follow that this is the only possible hypothesis or the best possible hypothesis. Hence the terms "**not reject**" or "**fail to reject**" are perhaps better than the term "**accept**."

Test for μ of the Normal Distribution with Known σ^2

The following example explains the three kinds of hypotheses.

EXAMPLE 2 Test for the Mean of the Normal Distribution with Known Variance

Let X be a normal random variable with variance $\sigma^2 = 9$. Using a sample of size $n = 10$ with mean $\bar{x}$, test the hypothesis $\mu = \mu_0 = 24$ against the three kinds of alternatives, namely,

$$\text{(a)}\quad \mu > \mu_0 \qquad \text{(b)}\quad \mu < \mu_0 \qquad \text{(c)}\quad \mu \neq \mu_0.$$

Solution. We choose the significance level $\alpha = 0.05$. An estimate of the mean will be obtained from

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n).$$

If the hypothesis is true, $\bar{X}$ is normal with mean $\mu = 24$ and variance $\sigma^2/n = 0.9$, see Theorem 1, Sec. 25.3. Hence we may obtain the critical value c from Table A8 in App. 5.

Case (a). Right-Sided Test. We determine c from $P(\bar{X} > c)_{\mu=24} = \alpha = 0.05$, that is,

$$P(\bar{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c - 24}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $(c - 24)/\sqrt{0.9} = 1.645$, and $c = 25.56$, which is greater than μ_0, as in the upper part of Fig. 533. If $\bar{x} \leqq 25.56$, the hypothesis is accepted. If $\bar{x} > 25.56$, it is rejected. The power function of the test is (Fig. 535)

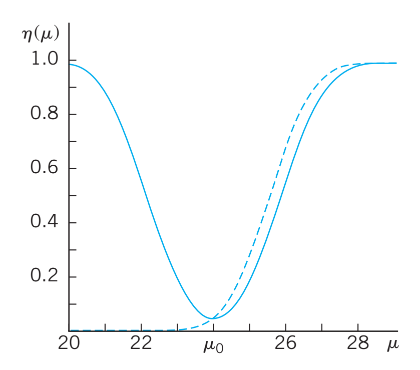


Fig. 535. Power function $\eta(\mu)$ in Example 2, case (a) (dashed) and case (c)

$$(7) \qquad \begin{aligned} \eta(\mu) &= P(\bar{X} > 25.56)_\mu = 1 - P(\bar{X} \leqq 25.56)_\mu \\ &= 1 - \Phi\left(\frac{25.56 - \mu}{\sqrt{0.9}}\right) = 1 - \Phi(26.94 - 1.05\mu) \end{aligned}$$

Case (b). Left-Sided Test. The critical value c is obtained from the equation

$$P(\bar{X} \leqq c)_{\mu=24} = \Phi\left(\frac{c - 24}{\sqrt{0.9}}\right) = \alpha = 0.05.$$

Table A8 in App. 5 yields $c = 24 - 1.56 = 22.44$. If $\bar{x} \geqq 22.44$, we accept the hypothesis. If $\bar{x} < 22.44$, we reject it. The power function of the test is

$$(8) \qquad \eta(\mu) = P(\bar{X} \leqq 22.44)_\mu = \Phi\left(\frac{22.44 - \mu}{\sqrt{0.9}}\right) = \Phi(23.65 - 1.05\mu).$$

Case (c). Two-Sided Test. Since the normal distribution is symmetric, we choose c_1 and c_2 equidistant from $\mu = 24$, say, $c_1 = 24 - k$ and $c_2 = 24 + k$, and determine k from

$$P(24 - k \leqq \bar{X} \leqq 24 + k)_{\mu=24} = \Phi\left(\frac{k}{\sqrt{0.9}}\right) - \Phi\left(-\frac{k}{\sqrt{0.9}}\right) = 1 - \alpha = 0.95.$$

Table A8 in App. 5 gives $k/\sqrt{0.9} = 1.960$, hence $k = 1.86$. This gives the values $c_1 = 24 - 1.86 = 22.14$ and $c_2 = 24 + 1.86 = 25.86$. If $\bar{x}$ is not smaller than c_1 and not greater than c_2, we accept the hypothesis. Otherwise we reject it. The power function of the test is (Fig. 535)

$$\eta(\mu) = P(\bar{X} < 22.14)_\mu + P(\bar{X} > 25.86)_\mu = P(\bar{X} < 22.14)_\mu + 1 - P(\bar{X} \leqq 25.86)_\mu$$

$$(9) \qquad = 1 + \Phi\left(\frac{22.14 - \mu}{\sqrt{0.9}}\right) - \Phi\left(\frac{25.86 - \mu}{\sqrt{0.9}}\right)$$

$$= 1 + \Phi(23.34 - 1.05\mu) - \Phi(27.26 - 1.05\mu).$$

Consequently, the operating characteristic $\beta(\mu) = 1 - \eta(\mu)$ (see before) is (Fig. 536)

$$\beta(\mu) = \Phi(27.26 - 1.05\mu) - \Phi(23.34 - 1.05\mu).$$

If we take a larger sample, say, of size $n = 100$ (instead of 10), then $\sigma^2/n = 0.09$ (instead of 0.9) and the critical values are $c_1 = 23.41$ and $c_2 = 24.59$, as can be readily verified. Then the operating characteristic of the test is

$$\beta(\mu) = \Phi\left(\frac{24.59 - \mu}{\sqrt{0.09}}\right) - \Phi\left(\frac{23.41 - \mu}{\sqrt{0.09}}\right)$$

$$= \Phi(81.97 - 3.33\mu) - \Phi(78.03 - 3.33\mu).$$

Figure 536 shows that the corresponding OC curve is steeper than that for $n = 10$. This means that the increase of n has led to an improvement of the test. In any practical case, n is chosen as small as possible but so large that the test brings out deviations between μ and μ_0 that are of practical interest. For instance, if deviations of ± 2 units are of interest, we see from Fig. 536 that $n = 10$ is much too small because when $\mu = 24 - 2 = 22$ or $\mu = 24 + 2 = 26$ β is almost 50%. On the other hand, we see that $n = 100$ is sufficient for that purpose. ■

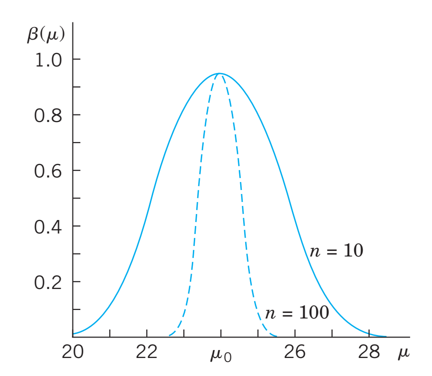


Fig. 536. Curves of the operating characteristic (OC curves) in Example 2, case (c), for two different sample sizes n

Test for μ When σ^2 Is Unknown, and for σ^2

EXAMPLE 3 Test for the Mean of the Normal Distribution with Unknown Variance

The tensile strength of a sample of $n = 16$ manila ropes (diameter 3 in.) was measured. The sample mean was $\bar{x} = 4482$ kg, and the sample standard deviation was $s = 115$ kg (N. C. Wiley, 41st Annual Meeting of the American Society for Testing Materials). Assuming that the tensile strength is a normal random variable, test the hypothesis $\mu_0 = 4500$ kg against the alternative $\mu_1 = 4400$ kg. Here μ_0 may be a value given by the manufacturer, while μ_1 may result from previous experience.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, it follows from Theorem 2 in Sec. 25.3, that the random variable

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} = \frac{\bar{X} - 4500}{S/4}$$

has a t-distribution with $n - 1 = 15$ d.f. The test is left-sided. The critical value c is obtained from $P(T < c)_{\mu_0} = \alpha = 0.05$. Table A9 in App. 5 gives $c = -1.75$. As an observed value of T we obtain from the sample $t = (4482 - 4500)/(115/4) = -0.626$. We see that $t > c$ and accept the hypothesis. For obtaining numeric values of the power of the test, we would need tables called noncentral Student t-tables; we shall not discuss this question here. ■

EXAMPLE 4 **Test for the Variance of the Normal Distribution**

Using a sample of size $n = 15$ and sample variance $s^2 = 13$ from a normal population, test the hypothesis $\sigma^2 = \sigma_0^2 = 10$ against the alternative $\sigma^2 = \sigma_1^2 = 20$.

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, then

$$Y = (n - 1)\frac{S^2}{\sigma_0^2} = 14\frac{S^2}{10} = 1.4S^2$$

has a chi-square distribution with $n - 1 = 14$ d.f. by Theorem 3, Sec. 25.3. From

$$P(Y > c) = \alpha = 0.05, \qquad \text{that is,} \qquad P(Y \leqq c) = 0.95,$$

and Table A10 in App. 5 with 14 degrees of freedom we obtain $c = 23.68$. This is the critical value of Y. Hence to $S^2 = \sigma_0^2 Y/(n - 1) = 0.714Y$ there corresponds the critical value $c^* = 0.714 \cdot 23.68 = 16.91$. Since $s^2 < c^*$, we accept the hypothesis.

If the alternative is true, the random variable $Y_1 = 14S^2/\sigma_1^2 = 0.7S^2$ has a chi-square distribution with 14 d.f. Hence our test has the power

$$\eta = P(S^2 > c^*)_{\sigma^2=20} = P(Y_1 > 0.7c^*)_{\sigma^2=20} = 1 - P(Y_1 \leqq 11.84)_{\sigma^2=20}.$$

From a more extensive table of the chi-square distribution (e.g. in Ref. [G3] or [G8]) or from your CAS, you see that $\eta \approx 62\%$. Hence the Type II risk is very large, namely, 38%. To make this risk smaller, we would have to increase the sample size. ■

Comparison of Means and Variances

EXAMPLE 5 **Comparison of the Means of Two Normal Distributions**

Using a sample $x_1, \cdots, x_{n_1}$ from a normal distribution with unknown mean μ_x and a sample $y_1, \cdots, y_{n_2}$ from another normal distribution with unknown mean μ_y, we want to test the hypothesis that the means are equal, $\mu_x = \mu_y$, against an alternative, say, $\mu_x > \mu_y$. The variances need not be known but are assumed to be equal.[3]

Two cases of comparing means are of practical importance:

Case A. *The samples have the* ***same size.*** *Furthermore, each value of the first sample corresponds to precisely one value of the other,* because corresponding values result from the same person or thing **(paired comparison)**—for example, two measurements of the same thing by two different methods or two measurements from the two eyes of the same person. More generally, they may result from pairs of *similar* individuals or things, for example, identical twins, pairs of used front tires from the same car, etc. Then we should form the differences of corresponding values and test the hypothesis that the population corresponding to the differences has mean 0, using the method in Example 3. If we have a choice, this method is better than the following.

[3]This assumption of equality of variances can be tested, as shown in the next example. If the test shows that they differ significantly, choose two samples of the same size $n_1 = n_2 = n$ (not too small, > 30, say), use the test in Example 2 together with the fact that (12) is an observed value of an approximately standardized normal random variable.

Case B. *The two samples are independent and not necessarily of the same size.* Then we may proceed as follows. Suppose that the alternative is $\mu_x > \mu_y$. We choose a significance level α. Then we compute the sample means $\bar{x}$ and $\bar{y}$ as well as $(n_1 - 1)s_x^2$ and $(n_2 - 1)s_y^2$, where s_x^2 and s_y^2 are the sample variances. Using Table A9 in App. 5 with $n_1 + n_2 - 2$ degrees of freedom, we now determine c from

$$P(T \leqq c) = 1 - \alpha. \tag{10}$$

We finally compute

$$t_0 = \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}}. \tag{11}$$

It can be shown that this is an observed value of a random variable that has a t-distribution with $n_1 + n_2 - 2$ degrees of freedom, provided the hypothesis is true. If $t_0 \leqq c$, the hypothesis is accepted. If $t_0 > c$, it is rejected.

If the alternative is $\mu_x \neq \mu_y$, then (10) must be replaced by

$$P(T \leqq c_1) = 0.5\alpha, \qquad P(T \leqq c_2) = 1 - 0.5\alpha. \tag{10*}$$

Note that for samples of equal size $n_1 = n_2 = n$, formula (11) reduces to

$$t_0 = \sqrt{n} \frac{\bar{x} - \bar{y}}{\sqrt{s_x^2 + s_y^2}}. \tag{12}$$

To illustrate the computations, let us consider the two samples $(x_1, \cdots, x_{n_1})$ and $(y_1, \cdots, y_{n_2})$ given by

105 108 86 103 103 107 124 105

and

89 92 84 97 103 107 111 97

showing the relative output of tin plate workers under two different working conditions [J. J. B. Worth, *Journal of Industrial Engineering* **9**, 249–253). Assuming that the corresponding populations are normal and have the same variance, let us test the hypothesis $\mu_x = \mu_y$ against the alternative $\mu_x \neq \mu_y$. (Equality of variances will be tested in the next example.)

Solution. We find

$$\bar{x} = 105.125, \qquad \bar{y} = 97.500, \qquad s_x^2 = 106.125. \qquad s_y^2 = 84.000.$$

We choose the significance level $\alpha = 5\%$. From (10*) with $0.5\alpha = 2.5\%$, $1 - 0.5\alpha = 97.5\%$ and Table A9 in App. 5 with 14 degrees of freedom we obtain $c_1 = -2.14$ and $c_2 = 2.14$. Formula (12) with $n = 8$ gives the value

$$t_0 = \sqrt{8} \cdot 7.625/\sqrt{190.125} = 1.56.$$

Since $c_1 \leqq t_0 \leqq c_2$, we ***accept the hypothesis*** $\mu_x = \mu_y$ that under both conditions the mean output is the same.

Case A applies to the example because the two first sample values correspond to a certain type of work, the next two were obtained in another kind of work, etc. So we may use the differences

16 16 2 6 0 0 13 8

of corresponding sample values and the method in Example 3 to test the hypothesis $\mu = 0$, where μ is the mean of the population corresponding to the differences. As a logical alternative we take $\mu \neq 0$. The sample mean is $\bar{d} = 7.625$, and the sample variance is $s^2 = 45.696$. Hence

$$t = \sqrt{8}\,(7.625 - 0)/\sqrt{45.696} = 3.19.$$

From $P(T \leqq c_1) = 2.5\%$, $P(T \leqq c_2) = 97.5\%$ and Table A9 in App. 5 with $n - 1 = 7$ degrees of freedom we obtain $c_1 = -2.36$, $c_2 = 2.36$ and ***reject the hypothesis*** because $t = 3.19$ does not lie between c_1 and c_2. Hence our present test, in which we used more information (but the same samples), shows that the difference in output is significant. ■

EXAMPLE 6 Comparison of the Variance of Two Normal Distributions

Using the two samples in the last example, test the hypothesis $\sigma_x^2 = \sigma_y^2$; assume that the corresponding populations are normal and the nature of the experiment suggests the alternative $\sigma_x^2 > \sigma_y^2$.

Solution. We find $s_x^2 = 106.125$, $s_y^2 = 84.000$. We choose the significance level $\alpha = 5\%$. Using $P(V \leqq c) = 1 - \alpha = 95\%$ and Table A11 in App. 5, with $(n_1 - 1, n_2 - 1) = (7, 7)$ degrees of freedom, we determine $c = 3.79$. We finally compute $v_0 = s_x^2/s_y^2 = 1.26$. Since $v_0 \leqq c$, we accept the hypothesis. If $v_0 < c$, we would reject it.

This test is justified by the fact that v_0 is an observed value of a random variable that has a so-called ***F*-distribution** with $(n_1 - 1, n_2 - 1)$ degrees of freedom, provided the hypothesis is true. (Proof in Ref. [G3] listed in App. 1.) The F-distribution with (m, n) degrees of freedom was introduced by R. A. Fisher[4] and has the distribution function $F(z) = 0$ if $z < 0$ and

$$F(z) = K_{mn} \int_0^z t^{(m-2)/2}(mt + n)^{-(m+n)/2}\, dt \qquad (z \geqq 0), \tag{13}$$

where $K_{mn} = m^{m/2}n^{n/2}\Gamma(\frac{1}{2}m + \frac{1}{2}n)/\Gamma(\frac{1}{2}m)\Gamma(\frac{1}{2}n)$. (For Γ see App. A3.1.) ■

This long section contained the basic ideas and concepts of testing, along with typical applications and you may perhaps want to review it quickly before going on, because the next sections concern an adaptation of these ideas to tasks of great practical importance and resulting tests in connection with quality control, acceptance (or rejection) of goods produced, and so on.

PROBLEM SET 25.4

1. From memory: Make a list of the three types of alternatives, each with a typical example of your own.

2. Make a list of methods in this section, each with the distribution needed in testing.

3. Test $\mu = 0$ against $\mu > 0$, assuming normality and using the sample 0, 1, -1, 3, -8, 6, 1 (deviations of the azimuth [multiples of 0.01 radian] in some revolution of a satellite). Choose $\alpha = 5\%$.

4. In one of his classical experiments Buffon obtained 2048 heads in tossing a coin 4040 times. Was the coin fair?

5. Do the same test as in Prob. 4, using a result by K. Pearson, who obtained 6019 heads in 12,000 trials.

6. Assuming normality and known variance $\sigma^2 = 9$, test the hypothesis $\mu = 60.0$ against the alternative $\mu = 57.0$ using a sample of size 20 with mean $\bar{x} = 58.50$ and choosing $\alpha = 5\%$.

7. How does the result in Prob. 6 change if we use a smaller sample, say, of size 5, the other data ($\bar{x} = 58.05$, $\alpha = 5\%$, etc.) remaining as before?

8. Determine the power of the test in Prob. 6.

9. What is the rejection region in Prob. 6 in the case of a two-sided test with $\alpha = 5\%$?

10. CAS EXPERIMENT. Tests of Means and Variances.
(a) Obtain 100 samples of size 10 each from the normal distribution with mean 100 and variance 25. For each sample, test the hypothesis $\mu_0 = 100$ against the alternative $\mu_1 > 100$ at the level of $\alpha = 10\%$. Record the number of rejections of the hypothesis. Do the whole experiment once more and compare.

(b) Set up a similar experiment for the variance of a normal distribution and perform it 100 times.

11. A firm sells oil in cans containing 5000 g oil per can and is interested to know whether the mean weight differs significantly from 5000 g at the 5% level, in which case the filling machine has to be adjusted. Set up a hypothesis and an alternative and perform the test, assuming normality and using a sample of 50 fillings with mean 4990 g and standard deviation 20 g.

---

[4]After the pioneering work of the English statistician and biologist, KARL PEARSON (1857–1936), the founder of the English school of statistics, and WILLIAM SEALY GOSSET (1876–1937), who discovered the t-distribution (and published under the name "Student"), the English statistician Sir RONALD AYLMER FISHER (1890–1962), professor of eugenics in London (1933–1943) and professor of genetics in Cambridge, England (1943–1957) and Adelaide, Australia (1957–1962), had great influence on the further development of modern statistics.

12. If a sample of 25 tires of a certain kind has a mean life of 37,000 miles and a standard deviation of 5000 miles, can the manufacturer claim that the true mean life of such tires is greater than 35,000 miles? Set up and test a corresponding hypothesis at the 5% level, assuming normality.

13. If simultaneous measurements of electric voltage by two different types of voltmeter yield the differences (in volts) 0.4, −0.6, 0.2, 0.0, 1.0, 1.4, 0.4, 1.6, can we assert at the 5% level that there is no significant difference in the calibration of the two types of instruments? Assume normality.

14. If a standard medication cures about 75% of patients with a certain disease and a new medication cured 310 of the first 400 patients on whom it was tried, can we conclude that the new medication is better? Choose $\alpha = 5\%$. First guess. Then calculate.

15. Suppose that in the past the standard deviation of weights of certain 100.0-oz packages filled by a machine was 0.8 oz. Test the hypothesis H_0: $\sigma = 0.8$ against the alternative H_1: $\sigma > 0.8$ (an undesirable increase), using a sample of 20 packages with standard deviation 1.0 oz and assuming normality. Choose $\alpha = 5\%$.

16. Suppose that in operating battery-powered electrical equipment, it is less expensive to replace all batteries at fixed intervals than to replace each battery individually when it breaks down, provided the standard deviation of the lifetime is less than a certain limit, say, less than 5 hours. Set up and apply a suitable test, using a sample of 28 values of lifetimes with standard deviation $s = 3.5$ hours and assuming normality: choose $\alpha = 5\%$.

17. Brand *A* gasoline was used in 16 similar automobiles under identical conditions. The corresponding sample of 16 values (miles per gallon) had mean 19.6 and standard deviation 0.4. Under the same conditions, high-power brand *B* gasoline gave a sample of 16 values with mean 20.2 and standard deviation 0.6. Is the mileage of *B* significantly better than that of *A*? Test at the 5% level; assume normality. First guess. Then calculate.

18. The two samples 70, 80, 30, 70, 60, 80 and 140, 120, 130, 120, 120, 130, 120 are values of the differences of temperatures (°C) of iron at two stages of casting, taken from two different crucibles. Is the variance of the first population larger than that of the second? Assume normality. Choose $\alpha = 5\%$.

19. Show that for a normal distribution the two types of errors in a test of a hypothesis H_0: $\mu = \mu_0$ against an alternative H_1: $\mu = \mu_1$ can be made as small as one pleases (not zero!) by taking the sample sufficiently large.

20. Test for equality of population means against the alternative that the means are different assuming normality, choosing $\alpha = 5\%$ and using two samples of sizes 12 and 18, with mean 10 and 14, respectively, and equal standard deviation 3.

25.5 Quality Control

The ideas on testing can be adapted and extended in various ways to serve basic practical needs in engineering and other fields. We show this in the remaining sections for some of the most important tasks solvable by statistical methods. As a first such area of problems, we discuss industrial quality control, a highly successful method used in various industries.

No production process is so perfect that all the products are completely alike. There is always a small variation that is caused by a great number of small, uncontrollable factors and must therefore be regarded as a chance variation. It is important to make sure that the products have required values (for example, length, strength, or whatever property may be essential in a particular case). For this purpose one makes a test of the hypothesis that the products have the required property, say, $\mu = \mu_0$, where μ_0 is a required value. If this is done after an entire lot has been produced (for example, a lot of 100,000 screws), the test will tell us how good or how bad the products are, but it it obviously too late to alter undesirable results. It is much better to test during the production run. This is done at regular intervals of time (for example, every hour or half-hour) and is called **quality control**. Each time a sample of the same size is taken, in practice 3 to 10 times. If the hypothesis is rejected, we stop the production and look for the cause of the trouble.

If we stop the production process even though it is progressing properly, we make a Type I error. If we do not stop the process even though something is not in order, we make a Type II error (see Sec. 25.4). The result of each test is marked in graphical form on what is called a **control chart**. This was proposed by W. A. Shewhart in 1924 and makes quality control particularly effective.

Control Chart for the Mean

An illustration and example of a control chart is given in the upper part of Fig. 537. This control chart for the mean shows the **lower control limit** LCL, the **center control line** CL, and the **upper control limit** UCL. The two **control limits** correspond to the critical values c_1 and c_2 in case (c) of Example 2 in Sec. 25.4. As soon as a sample mean falls outside the range between the control limits, we reject the hypothesis and assert that the

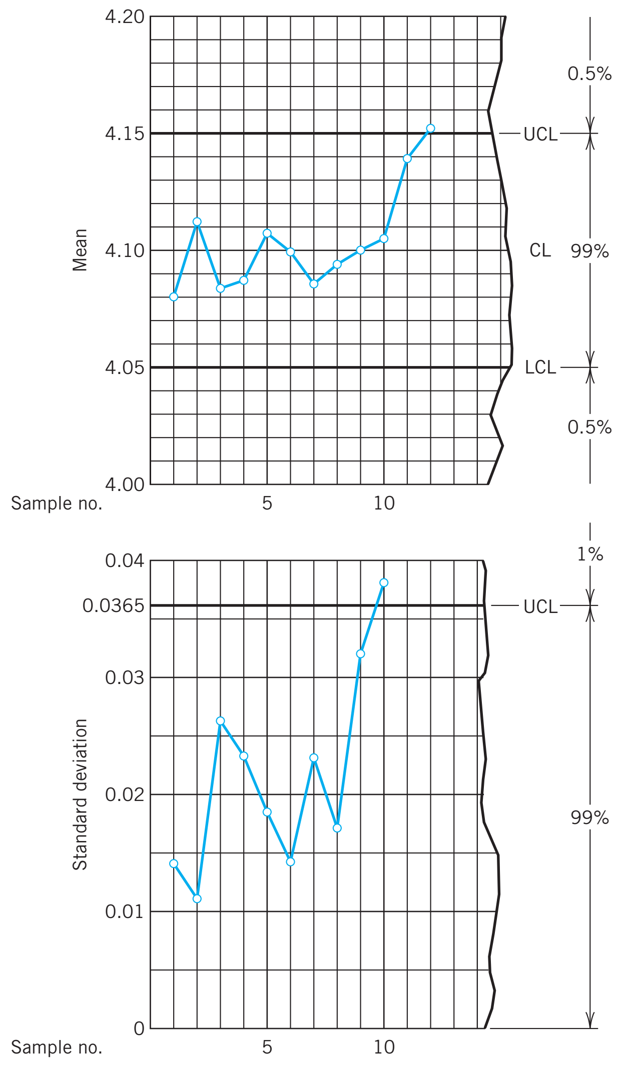


Fig. 537. Control charts for the mean (upper part of figure) and the standard deviation in the case of the samples on p. 1089

production process is "out of control"; that is, we assert that there has been a shift in process level. Action is called for whenever a point exceeds the limits.

If we choose control limits that are too loose, we shall not detect process shifts. On the other hand, if we choose control limits that are too tight, we shall be unable to run the process because of frequent searches for nonexistent trouble. The usual significance level is $\alpha = 1\%$. From Theorem 1 in Sec. 25.3 and Table A8 in App. 5 we see that in the case of the normal distribution the corresponding control limits for the mean are

$$\text{(1)} \qquad \text{LCL} = \mu_0 - 2.58\frac{\sigma}{\sqrt{n}}, \qquad \text{UCL} = \mu_0 + 2.58\frac{\sigma}{\sqrt{n}}.$$

Here σ is assumed to be known. If σ is unknown, we may compute the standard deviations of the first 20 or 30 samples and take their arithmetic mean as an approximation of σ. The broken line connecting the means in Fig. 537 is merely to display the results.

Additional, more subtle controls are often used in industry. For instance, one observes the motions of the sample means above and below the centerline, which should happen frequently. Accordingly, long runs (conventionally of length 7 or more) of means all above (or all below) the centerline could indicate trouble.

Table 25.5 Twelve Samples of Five Values Each (Diameter of Small Cylinders, Measured in Millimeters)

Sample Number	Sample Values					$\bar{x}$	s	R
1	4.06	4.08	4.08	4.08	4.10	4.080	0.014	0.04
2	4.10	4.10	4.12	4.12	4.12	4.112	0.011	0.02
3	4.06	4.06	4.08	4.10	4.12	4.084	0.026	0.06
4	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
5	4.08	4.10	4.12	4.12	4.12	4.108	0.018	0.04
6	4.08	4.10	4.10	4.10	4.12	4.100	0.014	0.04
7	4.06	4.08	4.08	4.10	4.12	4.088	0.023	0.06
8	4.08	4.08	4.10	4.10	4.12	4.096	0.017	0.04
9	4.06	4.08	4.10	4.12	4.14	4.100	0.032	0.08
10	4.06	4.08	4.10	4.12	4.16	4.104	0.038	0.10
11	4.12	4.14	4.14	4.14	4.16	4.140	0.014	0.04
12	4.14	4.14	4.16	4.16	4.16	4.152	0.011	0.02

Control Chart for the Variance

In addition to the mean, one often controls the variance, the standard deviation, or the range. To set up a control chart for the variance in the case of a normal distribution, we may employ the method in Example 4 of Sec. 25.4 for determining control limits. It is customary to use only one control limit, namely, an upper control limit. Now from Example 4 of Sec. 25.4 we have $S^2 = \sigma_0^2 Y/(n-1)$, where, because of our normality assumption, the random variable Y has a chi-square distribution with $n - 1$ degrees of freedom. Hence the desired control limit is

$$\text{(2)} \qquad \text{UCL} = \frac{\sigma^2 c}{n-1}$$

where c is obtained from the equation

$$P(Y > c) = \alpha, \qquad \text{that is,} \qquad P(Y \leqq c) = 1 - \alpha$$

and the table of the chi-square distribution (Table A10 in App. 5) with $n - 1$ degrees of freedom (or from your CAS); here α (5% or 1%, say) is the probability that in a properly running process an observed value s^2 of S^2 is greater than the upper control limit.

If we wanted a control chart for the variance with both an upper control limit UCL and a lower control limit LCL, these limits would be

$$\text{(3)} \qquad \text{LCL} = \frac{\sigma^2 c_1}{n-1} \qquad \text{and} \qquad \text{UCL} = \frac{\sigma^2 c_2}{n-1},$$

where c_1 and c_2 are obtained from Table A10 with $n - 1$ d.f. and the equations

$$\text{(4)} \qquad P(Y \leqq c_1) = \frac{\alpha}{2} \qquad \text{and} \qquad P(Y \leqq c_2) = 1 - \frac{\alpha}{2}.$$

Control Chart for the Standard Deviation

To set up a control chart for the standard deviation, we need an upper control limit

$$\text{(5)} \qquad \text{UCL} = \frac{\sigma\sqrt{c}}{\sqrt{n-1}}$$

obtained from (2). For example, in Table 25.5 we have $n = 5$. Assuming that the corresponding population is normal with standard deviation $\sigma = 0.02$ and choosing $\alpha = 1\%$, we obtain from the equation

$$P(Y \leqq c) = 1 - \alpha = 99\%$$

and Table A10 in App. 5 with 4 degrees of freedom the critical value $c = 13.28$ and from (5) the corresponding value

$$\text{UCL} = \frac{0.02\sqrt{13.28}}{\sqrt{4}} = 0.0365,$$

which is shown in the lower part of Fig. 537.

A control chart for the standard deviation with both an upper and a lower control limit is obtained from (3).

Control Chart for the Range

Instead of the variance or standard deviation, one often controls the **range** R (= largest sample value minus smallest sample value). It can be shown that in the case of the normal distribution, the standard deviation σ is proportional to the expectation of the random

variable R^* for which R is an observed value, say, $\sigma = \lambda_n E(R^*)$ where the factor of proportionality λ_n depends on the sample size n and has the values

n	2	3	4	5	6	7	8	9	10
$\lambda_n = \sigma/E(R^*)$	0.89	0.59	0.49	0.43	0.40	0.37	0.35	0.34	0.32

n	12	14	16	18	20	30	40	50
$\lambda_n = \sigma/E(R^*)$	0.31	0.29	0.28	0.28	0.27	0.25	0.23	0.22

Since R depends on two sample values only, it gives less information about a sample than s does. Clearly, the larger the sample size n is, the more information we lose in using R instead of s. A practical rule is to use s when n is larger than 10.

PROBLEM SET 25.5

1. Suppose a machine for filling cans with lubricating oil is set so that it will generate fillings which form a normal population with mean 1 gal and standard deviation 0.02 gal. Set up a control chart of the type shown in Fig. 537 for controlling the mean, that is, find LCL and UCL, assuming that the sample size is 4.

2. Three-sigma control chart. Show that in Prob. 1, the requirement of the significance level $\alpha = 0.3\%$ leads to LCL $= \mu - 3\sigma/\sqrt{n}$ and UCL $= \mu + 3\sigma/\sqrt{n}$, and find the corresponding numeric values.

3. What sample size should we choose in Prob. 1 if we want LCL and UCL somewhat closer together, say, UCL $-$ LCL $= 0.02$, without changing the significance level?

4. What effect on UCL $-$ LCL does it have if we double the sample size? If we switch from $\alpha = 1\%$ to $\alpha = 5\%$?

5. How should we change the sample size in controlling the mean of a normal population if we want UCL $-$ LCL to decrease to half its original value?

6. Graph the means of the following 10 samples (thickness of gaskets, coded values) on a control chart for means, assuming that the population is normal with mean 5 and standard deviation 1.16.

Time	10:00	11:00	12:00	13:00	14:00	15:00	16:00	17:00	18:00	19:00
	5	7	7	4	5	6	5	5	3	3
Sample	2	5	3	4	6	4	5	2	4	6
values	5	4	6	3	4	6	6	5	8	6
	6	4	5	6	6	4	4	3	4	8

7. Graph the ranges of the samples in Prob. 6 on a control chart for ranges.

8. Graph $\lambda_n = \sigma/E(R^*)$ as a function of n. Why is λ_n a monotone decreasing function of n?

9. Eight samples of size 2 were taken from a lot of screws. The values (length in inches) are

Sample No.	1	2	3	4	5	6	7	8
Length	3.50	3.51	3.49	3.52	3.53	3.49	3.48	3.52
	3.51	3.48	3.50	3.50	3.49	3.50	3.47	3.49

Assuming that the population is normal with mean 3.500 and variance 0.0004 and using (1), set up a control chart for the mean and graph the sample means on the chart.

10. Attribute control charts. Fifteen samples of size 100 were taken from a production of containers. The numbers of defectives (leaking containers) in those samples (in the order observed) were

1 4 5 4 9 7 0 5 6 13 0 2 1 12 8

From previous experience it was known that the average fraction defective is $p = 4\%$ provided that the process of production is running properly. Using the binomial distribution, set up a *fraction defective chart* (also called a ***p*-chart**), that is, choose the

LCL = 0 and determine the UCL for the fraction defective (in percent) by the use of 3-sigma limits, where σ^2 is the variance of the random variable

$$\bar{X} = \textit{Fraction defective in a sample of size } 100.$$

Is the process under control?

11. Number of defectives. Find formulas for the UCL, CL, and LCL (corresponding to 3σ-limits) in the case of a control chart for the number of defectives, assuming that, in a state of statistical control, the fraction of defectives is p.

12. CAS PROJECT. Control Charts. (a) Obtain 100 samples of 4 values each from the normal distribution with mean 8.0 and variance 0.16 and their means, variances, and ranges.

(b) Use these samples for making up a control chart for the mean.

(c) Use them on a control chart for the standard deviation.

(d) Make up a control chart for the range.

(e) Describe quantitative properties of the samples that you can see from those charts (e.g., whether the corresponding process is under control, whether the quantities observed vary randomly, etc.).

13. Since the presence of a point outside control limits for the mean indicates trouble, how often would we be making the mistake of looking for nonexistent trouble if we used **(a)** 1-sigma limits, **(b)** 2-sigma limits? Assume normality.

14. What LCL and UCL should we use instead of (1) if, instead of $\bar{x}$, we use the sum $x_1 + \cdots + x_n$ of the sample values? Determine these limits in the case of Fig. 537.

15. Number of defects per unit. A so-called *c-chart* or *defects-per-unit chart* is used for the control of the number X of defects per unit (for instance, the number of defects per 100 meters of paper, the number of missing rivets in an airplane wing, etc.). **(a)** Set up formulas for CL and LCL, UCL corresponding to $\mu \pm 3\sigma$, assuming that X has a Poisson distribution. **(b)** Compute CL, LCL, and UCL in a control process of the number of imperfections in sheet glass; assume that this number is 3.6 per sheet on the average when the process is in control.

25.6 Acceptance Sampling

Acceptance sampling is usually done when products leave the factory (or in some cases even within the factory). The standard situation in acceptance sampling is that a **producer** supplies to a **consumer** (a buyer or wholesaler) a lot of N items (a carton of screws, for instance). The decision to **accept** or **reject** the lot is made by determining the number x of **defectives** (= defective items) in a sample of size n from the lot. The lot is accepted if $x \leqq c$, where c is called the **acceptance number**, giving the allowable number of defectives. If $x > c$, the consumer rejects the lot. Clearly, producer and consumer must agree on a certain **sampling plan** giving n and c.

From the hypergeometric distribution we see that the event A: *"Accept the lot"* has probability (see Sec. 24.7)

$$P(A) = P(X \leqq c) = \sum_{x=0}^{c} \binom{M}{x} \binom{N-M}{n-x} \Big/ \binom{N}{n} \tag{1}$$

where M is the number of defectives in a lot of N items. In terms of the **fraction defective** $\theta = M/N$ we can write (1) as

$$P(A; \theta) = \sum_{x=0}^{c} \binom{N\theta}{x} \binom{N-N\theta}{n-x} \Big/ \binom{N}{n}. \tag{2}$$

$P(A; \theta)$ can assume $n + 1$ values corresponding to $\theta = 0, 1/N, 2/N, \cdots, N/N$; here, n and c are fixed. A monotone smooth curve through these points is called the **operating characteristic curve (OC curve)** of the sampling plan considered.

EXAMPLE 1 Sampling Plan

Suppose that certain tool bits are packaged 20 to a box, and the following sampling plan is used. A sample of two tool bits is drawn, and the corresponding box is accepted if and only if both bits in the sample are good. In this case, $N = 20$, $n = 2$, $c = 0$, and (2) takes the form (a factor 2 drops out)

$$P(A;\theta) = \binom{20\,\theta}{0}\binom{20 - 20\,\theta}{2}\Big/\binom{20}{2}$$

$$= \frac{(20 - 20\,\theta)(19 - 20\,\theta)}{380}.$$

The values of $P(A, \theta)$ for $\theta = 0, 1/20, 2/20, \cdots, 20/20$ and the resulting OC curve are shown in Fig. 538. (Verify!) ■

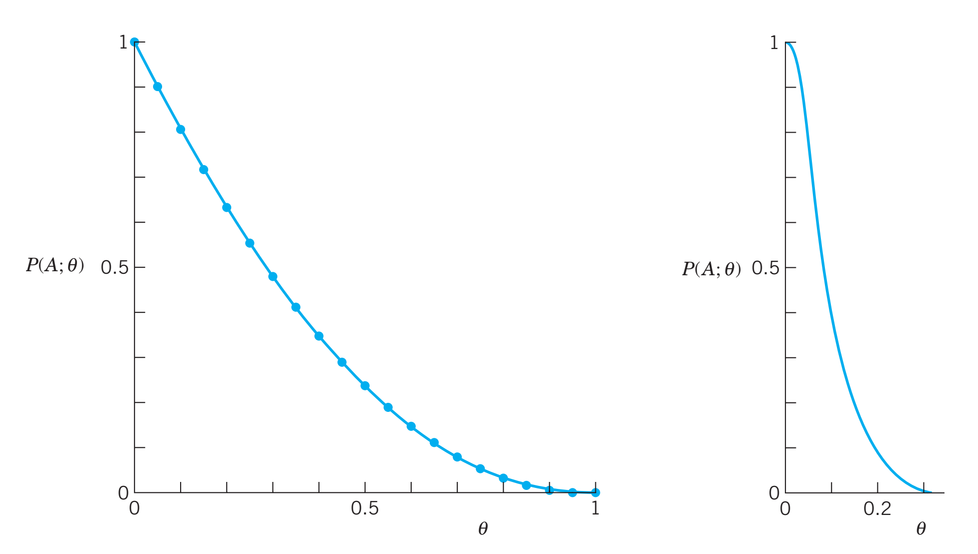


Fig. 538. OC curve of the sampling plan with $n = 2$ and $c = 0$ for lots of size $N = 20$

Fig. 539. OC curve in Example 2

In most practical cases θ will be small (less than 10%). Then if we take small samples compared to N, we can approximate (2) by the Poisson distribution (Sec. 24.7); thus

$$(3) \qquad P(A;\theta) \sim e^{-\mu} \sum_{x=0}^{c} \frac{\mu^x}{x!} \qquad (\mu = n\theta).$$

EXAMPLE 2 Sampling Plan. Poisson Distribution

Suppose that for large lots the following sampling plan is used. A sample of size $n = 20$ is taken. If it contains not more than one defective, the lot is accepted. If the sample contains two or more defectives, the lot is rejected. In this plan, we obtain from (3)

$$P(A;\theta) \sim e^{-20\,\theta}(1 + 20\,\theta),$$

The corresponding OC curve is shown in Fig. 539. ■

Errors in Acceptance Sampling

We show how acceptance sampling fits into general test theory (Sec. 25.4) and what this means from a practical point of view. The producer wants the probability α of rejecting

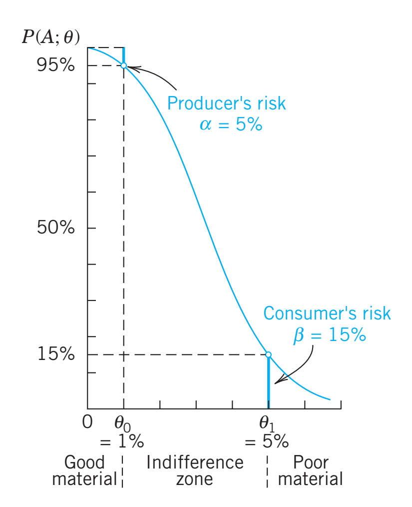


Fig. 540. OC curve, producer's and consumer's risks

an **acceptable lot** (a lot for which θ does not exceed a certain number θ_0 on which the two parties agree) to be small. θ_0 is called the **acceptable quality level** (AQL). Similarly, the consumer (the buyer) wants the probability β of accepting an **unacceptable lot** (a lot for which θ is greater than or equal to some θ_1) to be small. θ_1 is called the **lot tolerance percent defective** (LTPD) or the **rejectable quality level** (RQL). α is called **producer's risk**. It corresponds to a Type I error in Sec. 25.4. β is called **consumer's risk** and corresponds to a Type II error. Figure 540 shows an example. We see that the points $(\theta_0, 1-\alpha)$ and (θ_1, β) lie on the OC curve. It can be shown that for large lots we can choose $\theta_0, \theta_1\ (>\theta_0), \alpha, \beta$ and then determine n and c such that the OC curve runs very close to those prescribed points. Table 25.6 shows the analogy between acceptance sampling and hypothesis testing in Sec. 25.4.

Table 25.6 Acceptance Sampling and Hypothesis Testing

Acceptance Sampling	Hypothesis Testing
Acceptable quality level (AQL) $\theta = \theta_0$	Hypothesis $\theta = \theta_0$
Lot tolerance percent defectives (LTPD) $\theta = \theta_1$	Alternative $\theta = \theta_1$
Allowable number of defectives c	Critical value c
Producer's risk α of rejecting a lot with $\theta \leqq \theta_0$	Probability α of making a Type I error (significance level)
Consumer's risk β of accepting a lot with $\theta \geqq \theta_1$	Probability β of making a Type II error

Rectification

Rectification of a *rejected* lot means that the lot is inspected item by item and all defectives are removed and replaced by nondefective items. (This may be too expensive if the lot is cheap; in this case the lot may be sold at a cut-rate price or scrapped.) If a production turns out $100\theta\%$ defectives, then in K lots of size N each, $KN\theta$ of the KN items are

defectives. Now $KP(A; \theta)$ of these lots are accepted. These contain $KPN\theta$ defectives, whereas the rejected and rectified lots contain no defectives, because of the rectification. Hence after the rectification the fraction defective in all K lots equals $KPN\theta/KN$. This is called the **average outgoing quality** (AOQ); thus

$$\text{AOQ}(\theta) = \theta P(A; \theta). \qquad (4)$$

Figure 541 shows an example. Since $\text{AOQ}(0) = 0$ and $P(A; 1) = 0$, the AOQ curve has a maximum at some $\theta = \theta^*$, giving the **average outgoing quality limit** (AOQL). This is the worst average quality that may be expected to be accepted under rectification.

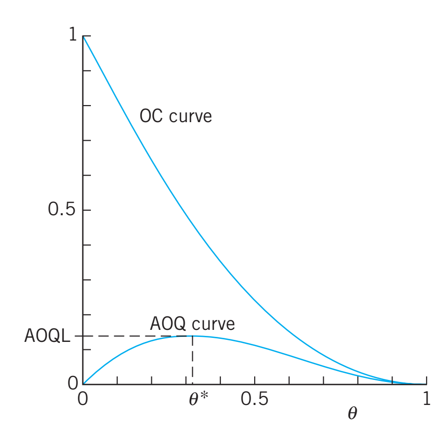


Fig. 541. OC curve and AOQ curve for the sampling plan in Fig. 538

PROBLEM SET 25.6

1. Lots of kitchen knives are inspected by a sampling plan that uses a sample of size 20 and the acceptance number $c = 1$. What is the probability of accepting a lot with $1\%, 2\%, 10\%$ defectives (knives with dull blades)? Use Table A6 of the Poisson distribution in App. 5. Graph the OC curve.
2. What happens in Prob. 1 if the sample size is increased to 50? First guess. Then calculate. Graph the OC curve and compare.
3. How will the probabilities in Prob. 1 with $n = 20$ change (up or down) if we decrease c to zero? First guess.
4. What are the producer's and consumer's risks in Prob. 1 if the AQL is 2% and the RQL is 15%?
5. Lots of copper pipes are inspected according to a sample plan that uses sample size 25 and acceptance number 1. Graph the OC curve of the plan, using the Poisson approximation. Find the producer's risk if the AQL is 1.5%.
6. Graph the AOQ curve in Prob. 5. Determine the AOQL, assuming that rectification is applied.
7. In Example 1 in the text, what are the producer's and consumer's risks if the AQL is 0.1 and the RQL is 0.6?
8. What happens in Example 1 in the text if we increase the sample size to $n = 3$, leaving the other data as before? Compute $P(A; 0.1)$ and $P(A; 0.2)$ and compare with Example 1.
9. Graph and compare sampling plans with $c = 1$ and increasing values of n, say, $n = 2, 3, 4$. (Use the binomial distribution.)
10. Find the binomial approximation of the hypergeometric distribution in Example 1 in the text and compare the approximate and the accurate values.

11. Samples of 3 fuses are drawn from lots and a lot is accepted if in the corresponding sample we find no more than 1 defective fuse. Criticize this sampling plan. In particular, find the probability of accepting a lot that is 50% defective. (Use the binomial distribution (7), Sec. 24.7.)

12. If in a sampling plan for large lots of spark plugs, the sample size is 100 and we want the AQL to be 5% and the producer's risk 2%, what acceptance number c should we choose? (Use the normal approximation of the binomial distribution in Sec. 24.8.)

13. What is the consumer's risk in Prob. 12 if we want the RQL to be 12%? Use $c = 9$ from the answer of Prob. 12.

14. A lot of batteries for wrist watches is accepted if and only if a sample of 20 contains at most 1 defective. Graph the OC and AOQ curves. Find AOQL. [Use (3).]

15. Graph the OC curve and the AOQ curve for the single sampling plan for large lots with $n = 5$ and $c = 0$, and find the AOQL.

25.7 Goodness of Fit. χ^2-Test

To test for **goodness of fit** means that we wish to test that a certain function $F(x)$ is the distribution function of a distribution from which we have a sample $x_1, \cdots, x_n$. Then we test whether the **sample distribution function** $\tilde{F}(x)$ defined by

$$\tilde{F}(x) = \textit{Sum of the relative frequencies of all sample values } x_j \textit{ not exceeding } x$$

fits $F(x)$ "sufficiently well." If this is so, we shall accept the hypothesis that $F(x)$ is the distribution function of the population; if not, we shall reject the hypothesis.

This test is of considerable practical importance, and it differs in character from the tests for parameters (μ, σ^2, etc.) considered so far.

To test in that fashion, we have to know how much $\tilde{F}(x)$ can differ from $F(x)$ if the hypothesis is true. Hence we must first introduce a quantity that measures the deviation of $\tilde{F}(x)$ from $F(x)$, and we must know the probability distribution of this quantity under the assumption that the hypothesis is true. Then we proceed as follows. We determine a number c such that, if the hypothesis is true, a deviation greater than c has a small preassigned probability. If, nevertheless, a deviation greater than c occurs, we have reason to doubt that the hypothesis is true and we reject it. On the other hand, if the deviation does not exceed c, so that $\tilde{F}(x)$ approximates $F(x)$ sufficiently well, we accept the hypothesis. Of course, if we accept the hypothesis, this means that we have insufficient evidence to reject it, and this does not exclude the possibility that there are other functions that would not be rejected in the test. In this respect the situation is quite similar to that in Sec. 25.4.

Table 25.7 shows a test of that type, which was introduced by R. A. Fisher. This test is justified by the fact that if the hypothesis is true, then χ_0^2 is an observed value of a random variable whose distribution function approaches that of the chi-square distribution with $K - 1$ degrees of freedom (or $K - r - 1$ degrees of freedom if r parameters are estimated) as n approaches infinity. The requirement that at least five sample values lie in each interval in Table 25.7 results from the fact that for finite n that random variable has only *approximately* a chi-square distribution. A proof can be found in Ref. [G3] listed in App. 1. If the sample is so small that the requirement cannot be satisfied, one may continue with the test, but then use the result with caution.

Table 25.7 Chi-square Test for the Hypothesis That $F(x)$ is the Distribution Function of a Population from Which a Sample $x_1, \cdots, x_n$ is Taken

Step 1. Subdivide the x-axis into K intervals $I_1, I_2, \cdots, I_K$ such that each interval contains at least 5 values of the given sample $x_1, \cdots, x_n$. Determine the number b_j of sample values in the interval I_j, where $j = 1, \cdots, K$. If a sample value lies at a common boundary point of two intervals, add 0.5 to each of the two corresponding b_j.

Step 2. Using $F(x)$, compute the probability p_j that the random variable X under consideration assumes any value in the interval I_j, where $j = 1, \cdots, K$. Compute

$$e_j = np_j.$$

(This is the number of sample values theoretically expected in I_j if the hypothesis is true.)

Step 3. Compute the deviation

$$\chi_0^2 = \sum_{j=1}^{K} \frac{(b_j - e_j)^2}{e_j}. \tag{1}$$

Step 4. Choose a significance level (5%, 1%, or the like).

Step 5. Determine the solution c of the equation

$$P(\chi^2 \leqq c) = 1 - \alpha$$

from the table of the chi-sqare distribution with $K - 1$ degrees of freedom (Table A10 in App. 5). If r parameters of $F(x)$ are unknown and their maximum likelihood estimates (Sec. 25.2) are used, then use $K - r - 1$ degrees of freedom (instead of $K - 1$). If $\chi_0^2 \leqq c$, accept the hypothesis. If $\chi_0^2 > c$, reject the hypothesis.

Table 25.8 Sample of 100 Values of the Splitting Tensile Strength (lb/in.^{2}) of Concrete Cylinders

320	380	340	410	380	340	360	350	320	370
350	340	350	360	370	350	380	370	300	420
370	390	390	440	330	390	330	360	400	370
320	350	360	340	340	350	350	390	380	340
400	360	350	390	400	350	360	340	370	420
420	400	350	370	330	320	390	380	400	370
390	330	360	380	350	330	360	300	360	360
360	390	350	370	370	350	390	370	370	340
370	400	360	350	380	380	360	340	330	370
340	360	390	400	370	410	360	400	340	360

D. L. IVEY, Splitting tensile tests on structural lightweight aggregate concrete. Texas Transportation Institute, College Station, Texas.

EXAMPLE 1 Test of Normality

Test whether the population from which the sample in Table 25.8 was taken is normal.

Solution. Table 25.8 shows the values (column by column) in the order obtained in the experiment. Table 25.9 gives the frequency distribution and Fig. 542 the histogram. It is hard to guess the outcome of the test—does the histogram resemble a normal density curve sufficiently well or not?

The maximum likelihood estimates for μ and σ^2 are $\hat{\mu} = \bar{x} = 364.7$ and $\tilde{\sigma}^2 = 712.9$. The computation in Table 25.10 yields $\chi_0^2 = 2.688$. It is very interesting that the interval $375 \cdots 385$ contributes over 50% of χ_0^2. From the histogram we see that the corresponding frequency looks much too small. The second largest contribution comes from $395 \cdots 405$, and the histogram shows that the frequency seems somewhat too large, which is perhaps not obvious from inspection.

Table 25.9 Frequency Table of the Sample in Table 25.8

1 Tensile Strength x [lb/in.^{2}]	2 Absolute Frequency	3 Relative Frequency $\tilde{f}(x)$	4 Cumulative Absolute Frequency	5 Cumulative Relative Frequency $\tilde{F}(x)$
300	2	0.02	2	0.02
310	0	0.00	2	0.02
320	4	0.04	6	0.06
330	6	0.06	12	0.12
340	11	0.11	23	0.23
350	14	0.14	37	0.37
360	16	0.16	53	0.53
370	15	0.15	68	0.68
380	8	0.08	76	0.76
390	10	0.10	86	0.86
400	8	0.08	94	0.94
410	2	0.02	96	0.96
420	3	0.03	99	0.99
430	0	0.00	99	0.99
440	1	0.01	100	1.00

We choose $\alpha = 5\%$. Since $K = 10$ and we estimated $r = 2$ parameters we have to use Table A10 in App. 5 with $K - r - 1 = 7$ degrees of freedom. We find $c = 14.07$ as the solution of $P(\chi^2 \leqq c) = 95\%$. Since $\chi_0^2 < c$, we accept the hypothesis that the population is normal. ■

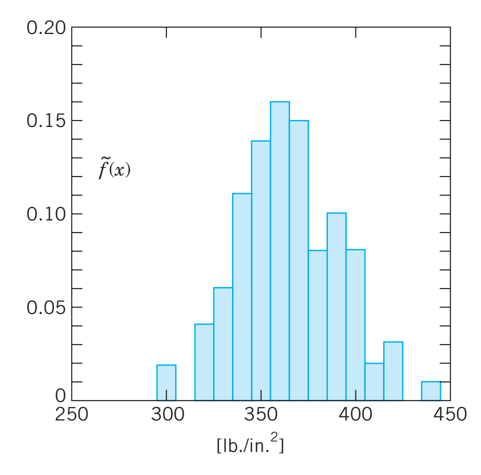


Fig. 542. Frequency histogram of the sample in Table 25.8

Table 25.10 Computations in Example 1

x_j	$\frac{x_j - 364.7}{26.7}$	$\Phi\left(\frac{x_j - 364.7}{26.7}\right)$	e_j	b_j	Term in (1)
$-\infty \cdots 325$	$-\infty \cdots -1.49$	$0.0000 \cdots 0.0681$	6.81	6	0.096
$325 \cdots 335$	$-1.49 \cdots -1.11$	$0.0681 \cdots 0.1335$	6.54	6	0.045
$335 \cdots 345$	$-1.11 \cdots -0.74$	$0.1335 \cdots 0.2296$	9.61	11	0.201
$345 \cdots 355$	$-0.74 \cdots -0.36$	$0.2296 \cdots 0.3594$	12.98	14	0.080
$355 \cdots 365$	$-0.36 \cdots 0.01$	$0.3594 \cdots 0.5040$	14.46	16	0.164
$365 \cdots 375$	$0.01 \cdots 0.39$	$0.5040 \cdots 0.6517$	14.77	15	0.0004
$375 \cdots 385$	$0.39 \cdots 0.76$	$0.6517 \cdots 0.7764$	12.47	8	1.602
$385 \cdots 395$	$0.76 \cdots 1.13$	$0.7764 \cdots 0.8708$	9.44	10	0.033
$395 \cdots 405$	$1.13 \cdots 1.51$	$0.8708 \cdots 0.9345$	6.37	8	0.417
$405 \cdots \infty$	$1.51 \cdots \infty$	$0.9345 \cdots 1.0000$	6.55	6	0.046

$\chi_0^2 = 2.688$

PROBLEM SET 25.7

1. Verify the calculations in Example 1 of the text.
2. If it is known that 25% of certain steel rods produced by a standard process will break when subjected to a load of 5000 lb, can we claim that a new, less expensive process yields the same breakage rate if we find that in a sample of 80 rods produced by the new process, 27 rods broke when subjected to that load? (Use $\alpha = 5\%$.)
3. If 100 flips of a coin result in 40 heads and 60 tails, can we assert on the 5% level that the coin is fair?
4. If in 10 flips of a coin we get the same ratio as in Prob. 3 (4 heads and 6 tails), is the conclusion the same as in Prob. 3? First conjecture, then compute.
5. Can you claim, on a 5% level, that a die is fair if 60 trials give $1, \cdots, 6$ with absolute frequencies 10, 13, 9, 11, 9, 8?
6. Solve Prob. 5 if rolling a die 180 times gives 33, 27, 29, 35, 25, 31.
7. If a service station had served 60, 49, 56, 46, 68, 39 cars from Monday through Friday between 1 P.M. and 2 P.M., can one claim on a 5% level that the differences are due to randomness? First guess. Then calculate.
8. A manufacturer claims that in a process of producing drill bits, only 2.5% of the bits are dull. Test the claim against the alternative that more than 2.5% of the bits are dull, using a sample of 400 bits containing 17 dull ones. Use $\alpha = 5\%$.
9. In a table of properly rounded function values, even and odd last decimals should appear about equally often. Test this for the 90 values of $J_1(x)$ in Table A1 in App. 5.
10. **TEAM PROJECT. Difficulty with Random Selection.** 77 students were asked to choose 3 of the integers $11, 12, 13, \cdots, 30$ completely arbitrarily. The amazing result was as follows.

Number	11	12	13	14	15	16	17	18	19	20
Frequ.	11	10	20	8	13	9	21	9	16	8

Number	21	22	23	24	25	26	27	28	29	30
Frequ.	12	8	15	10	10	9	12	8	13	9

If the selection were completely random, the following hypotheses should be true.
(a) The 20 numbers are equally likely.
(b) The 10 even numbers together are as likely as the 10 odd numbers together.
(c) The 6 prime numbers together have probability 0.3 and the 14 other numbers together have probability 0.7.
Test these hypotheses, using $\alpha = 5\%$. Design further experiments that illustrate the difficulties of random selection.

11. **CAS EXPERIMENT. Random Number Generator.** Check your generator experimentally by imitating results of n trials of rolling a fair die, with a convenient n (e.g., 60 or 300 or the like). Do this many times and see whether you can notice any "nonrandomness" features, for example, too few Sixes, too many even numbers, etc., or whether your generator seems to work properly. Design and perform other kinds of checks.
12. Test for normality at the 1% level using a sample of $n = 79$ (rounded) values x (tensile strength [kg/mm^2]

of steel sheets of 0.3 mm thickness). $a = a(x) =$ absolute frequency. (Take the first two values together, also the last three, to get $K = 5$.)

x	57	58	59	60	61	62	63	64
a	4	10	17	27	8	9	3	1

13. Mendel's pathbreaking experiments. In a famous plant-crossing experiment, the Austrian Augustinian father Gregor Mendel (1822–1884) obtained 355 yellow and 123 green peas. Test whether this agrees with Mendel's theory according to which the ratio should be 3:1.

14. Accidents in a foundry. Does the random variable X = *Number of accidents per week* have a Poisson distribution if, within 50 weeks, 33 were accident-free, 1 accident occurred in 11 of the 50 weeks, 2 in 6 of the weeks, and more than 2 accidents in no week? Choose $\alpha = 5\%$.

15. Radioactivity. Rutherford-Geiger experiments. Using the given sample, test that the corresponding population has a Poisson distribution. x is the number of alpha particles per 7.5-s intervals observed by E. Rutherford and H. Geiger in one of their classical experiments in 1910, and $a(x)$ is the absolute frequency (= number of time periods during which exactly x particles were observed). Use $\alpha = 5\%$.

x	0	1	2	3	4	5	6
a	57	203	383	525	532	408	273
x	7	8	9	10	11	12	$\geqq 13$
a	139	45	27	10	4	2	0

25.8 Nonparametric Tests

Nonparametric tests, also called **distribution-free tests**, are valid for any distribution. Hence they are used in cases when the kind of distribution is unknown, or is known but such that no tests specifically designed for it are available. In this section we shall explain the basic idea of these tests, which are based on "**order statistics**" and are rather simple. If there is a choice, then tests designed for a specific distribution generally give better results than do nonparametric tests. For instance, this applies to the tests in Sec. 25.4 for the normal distribution.

We shall discuss two tests in terms of typical examples. In deriving the distributions used in the test, it is essential that the distributions, from which we sample, are continuous. (Nonparametric tests can also be derived for discrete distributions, but this is slightly more complicated.)

EXAMPLE 1 Sign Test for the Median

A **median** of the population is a solution $x = \tilde{\mu}$ of the equation $F(x) = 0.5$, where F is the distribution function of the population.

Suppose that eight radio operators were tested, first in rooms without air-conditioning and then in air-conditioned rooms over the same period of time, and the difference of errors (unconditioned minus conditioned) were

$$9 \quad 4 \quad 0 \quad 6 \quad 4 \quad 0 \quad 7 \quad 11.$$

Test the hypothesis $\tilde{\mu} = 0$ (that is, air-conditioning has no effect) against the alternative $\overline{\mu} > 0$ (that is, inferior performance in unconditioned rooms).

Solution. We choose the significance level $\alpha = 5\%$. If the hypothesis is true, the probability p of a positive difference is the same as that of a negative difference. Hence in this case, $p = 0.5$, and the random variable

$$X = \textit{Number of positive values among n values}$$

has a binomial distribution with $p = 0.5$. Our sample has eight values. We omit the values 0, which do not contribute to the decision. Then six values are left, all of which are positive. Since

$$\begin{aligned} P(X = 6) &= \binom{6}{6}(0.5)^6(0.5)^0 \\ &= 0.0156 \\ &= 1.56\% \end{aligned}$$

we have observed an event whose probability is very small if the hypothesis is true; in fact $1.56\% < \alpha = 5\%$. Hence we assert that the alternative $\tilde{\mu} > 0$ is true. That is, the number of errors made in unconditioned rooms is significantly higher, so that installation of air conditioning should be considered. ■

EXAMPLE 2 Test for Arbitrary Trend

A certain machine is used for cutting lengths of wire. Five successive pieces had the lengths

$$29 \quad 31 \quad 28 \quad 30 \quad 32.$$

Using this sample, test the hypothesis that there is **no trend**, that is, the machine does not have the tendency to produce longer and longer pieces or shorter and shorter pieces. Assume that the type of machine suggests the alternative that there is *positive trend,* that is, there is the tendency of successive pieces to get longer.

Solution. We count the number of **transpositions** in the sample, that is, the number of times a larger value precedes a smaller value:

29 precedes 28 (1 transposition),

31 precedes 28 and 30 (2 transpositions).

The remaining three sample values follow in ascending order. Hence in the sample there are $1 + 2 = 3$ transpositions. We now consider the random variable

$$T = \textit{Number of transpositions.}$$

If the hypothesis is true (no trend), then each of the $5! = 120$ permutations of five elements 1 2 3 4 5 has the same probability (1/120). We arrange these permutations according to their number of transpositions:

$T = 0$	$T = 1$	$T = 2$	$T = 3$
1 2 3 4 5	1 2 3 5 4	1 2 4 5 3	1 2 5 4 3
	1 2 4 3 5	1 2 5 3 4	1 3 4 5 2
	1 3 2 4 5	1 3 2 5 4	1 3 5 2 4
	2 1 3 4 5	1 3 4 2 5	1 4 2 5 3
		1 4 2 3 5	1 4 3 2 5
		2 1 3 5 4	1 5 2 3 4
		2 1 4 3 5	2 1 4 5 3
		2 3 1 4 5	2 1 5 3 4 etc.
		3 1 2 4 5	2 3 1 5 4
			2 3 4 1 5
			2 4 1 3 5
			3 1 2 5 4
			3 1 4 2 5
			3 2 1 4 5
			4 1 2 3 5

From this we obtain

$$P(T \leqq 3) = \tfrac{1}{120} + \tfrac{4}{120} + \tfrac{9}{120} + \tfrac{15}{120} = \tfrac{29}{120} = 24\%.$$

We accept the hypothesis because we have observed an event that has a relatively large probability (certainly much more than 5%) if the hypothesis is true.

Values of the distribution function of T in the case of no trend are shown in Table A12, App. 5. For instance, if $n = 3$, then $F(0) = 0.167$, $F(1) = 0.500$, $F(2) = 1 - 0.167$. If $n = 4$, then $F(0) = 0.042$, $F(1) = 0.167$, $F(2) = 0.375$, $F(3) = 1 - 0.375$, $F(4) = 1 - 0.167$, and so on.

Our method and those values refer to *continuous* distributions. Theoretically, we may then expect that all the values of a sample are different. Practically, some sample values may still be equal, because of rounding: If m values are equal, add $m(m-1)/4$ (= mean value of the transpositions in the case of the permutations of m elements), that is, $\frac{1}{2}$ for each pair of equal values, $\frac{3}{2}$ for each triple, etc. ■

PROBLEM SET 25.8

1. What would change in Example 1 had we observed only 5 positive values? Only 4?

2. Test $\tilde{\mu} = 0$ against $\tilde{\mu} > 0$, using $1, -1, 1, 3, -8, 6, 0$ (deviations of the azimuth [multiples of 0.01 radian] in some revolution of a satellite).

3. Are oil filters of type A better than type B filters if in 11 trials, A gave cleaner oil than B in 7 cases, B gave cleaner oil than A in 1 case, whereas in 3 of the trials the results for A and B were practically the same?

4. Does a process of producing stainless steel pipes of length 20 ft for nuclear reactors need adjustment if, in a sample, 4 pipes have the exact length and 15 are shorter and 3 longer than 20 ft? Use the normal approximation of the binomial distribution.

5. Do the computations in Prob. 4 without the use of the DeMoivre–Laplace limit theorem in Sec. 24.8.

6. Thirty new employees were grouped into 15 pairs of similar intelligence and experience and were then instructed in data processing by an old method (A) applied to one (randomly selected) person of each pair, and by a new presumably better method (B) applied to the other person of each pair. Test for equality of methods against the alternative that (B) is better than (A), using the following scores obtained after the end of the training period.

A	60	70	80	85	75	40	70	45	95	80	90	60	80	75	65
B	65	85	85	80	95	65	100	60	90	85	100	75	90	60	80

7. Assuming normality, solve Prob. 6 by a suitable test from Sec. 25.4.

8. In a clinical experiment, each of 10 patients were given two different sedatives A and B. The following table shows the effect (increase of sleeping time, measured in hours). Using the sign test, find out whether the difference is significant.

A	1.9	0.8	1.1	0.1	−0.1	4.4	5.5	1.6	4.6	3.4
B	0.7	−1.6	−0.2	−1.2	−0.1	3.4	3.7	0.8	0.0	2.0
Difference	1.2	2.4	1.3	1.3	0.0	1.0	1.8	0.8	4.6	1.4

9. Assuming that the populations corresponding to the samples in Prob. 8 are normal, apply a suitable test for the normal distribution.

10. Test whether a thermostatic switch is properly set to 50°C against the alternative that its setting is too low. Use a sample of 9 values, 8 of which are less than 50°C and 1 is greater.

11. How would you proceed in the sign test if the hypothesis is $\tilde{\mu} = \tilde{\mu}_0$ (any number) instead of $\tilde{\mu} = 0$?

12. Test the hypothesis that, for a certain type of voltmeter, readings are independent of temperature T [°C] against the alternative that they tend to increase with T. Use a sample of values obtained by applying a constant voltage:

Temperature T [°C]	10	20	30	40	50
Reading V [volts]	99.5	101.1	100.4	100.8	101.6

13. Does the amount of fertilizer increase the yield of wheat X [kg/plot]? Use a sample of values ordered according to increasing amounts of fertilizer:

33.4 35.3 31.6 35.0 36.1 37.6 36.5 38.7.

14. Apply the test explained in Example 2 to the following data (x = diastolic blood pressure [mm Hg], y = weight of heart [in grams] of 10 patients who died of cerebral hemorrhage).

x	121	120	95	123	140	112	92	100	102	91
y	521	465	352	455	490	388	301	395	375	418

15. Does an increase in temperature cause an increase of the yield of a chemical reaction from which the following sample was taken?

Temperature [°C]	10	20	30	40	60	80
Yield [kg/min]	0.6	1.1	0.9	1.6	1.2	2.0

25.9 Regression. Fitting Straight Lines. Correlation

So far we were concerned with random experiments in which we observed a single quantity (random variable) and got samples whose values were single numbers. In this section we discuss experiments in which we observe or measure two quantities simultaneously, so that we get samples of *pairs* of values $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$. Most applications involve one of two kinds of experiments, as follows.

1. In **regression analysis** one of the two variables, call it x, can be regarded as an ordinary variable because we can measure it without substantial error or we can even give it values we want. x is called the **independent variable**, or sometimes the **controlled variable** because we can control it (set it at values we choose). The other variable, Y, is a random variable, and we are interested in the dependence of Y on x. Typical examples are the dependence of the blood pressure Y on the age x of a person or, as we shall now say, the regression of Y on x, the regression of the gain of weight Y of certain animals on the daily ration of food x, the regression of the heat conductivity Y of cork on the specific weight x of the cork, etc.
2. In **correlation analysis** both quantities are random variables and we are interested in relations between them. Examples are the relation (one says "correlation") between wear X and wear Y of the front tires of cars, between grades X and Y of students in mathematics and in physics, respectively, between the hardness X of steel plates in the center and the hardness Y near the edges of the plates, etc.

Regression Analysis

In regression analysis the dependence of Y on x is a dependence of the mean μ of Y on x, so that $\mu = \mu(x)$ is a function in the ordinary sense. The curve of $\mu(x)$ is called the **regression curve** of Y on x.

In this section we discuss the simplest case, namely, that of a straight **regression line**

(1)
$$\mu(x) = \kappa_0 + \kappa_1 x.$$

Then we may want to graph the sample values as n points in the xY-plane, fit a straight line through them, and use it for estimating $\mu(x)$ at values of x that interest us, so that we know what values of Y we can expect for those x. Fitting that line by eye would not be good because it would be subjective; that is, different persons' results would come out differently, particularly if the points are scattered. So we need a mathematical method that gives a unique result depending only on the n points. A widely used procedure is the method of least squares by Gauss and Legendre. For our task we may formulate it as follows.

Least Squares Principle

The straight line should be fitted through the given points so that the sum of the squares of the distances of those points from the straight line is minimum, where the distance is measured in the vertical direction (the y-direction). (Formulas below.)

To get uniqueness of the straight line, we need some extra condition. To see this, take the sample (0, 1), (0, −1). Then all the lines $y = k_1 x$ with any k_1 satisfy the principle. (Can you see it?) The following assumption will imply uniqueness, as we shall find out.

General Assumption (A1)

The x-values $x_1, \cdots, x_n$ in our sample $(x_1, y_1), \cdots, (x_n, y_n)$ are not all equal.

From a given sample $(x_1, y_1), \cdots, (x_n, y_n)$ we shall now determine a straight line by least squares. We write the line as

$$y = k_0 + k_1 x \tag{2}$$

and call it the **sample regression line** because it will be the counterpart of the population regression line (1).

Now a sample point (x_j, y_j) has the vertical distance (distance measured in the y-direction) from (2) given by

$$|y_j - (k_0 + k_1 x_j)| \qquad \text{(see Fig. 543).}$$

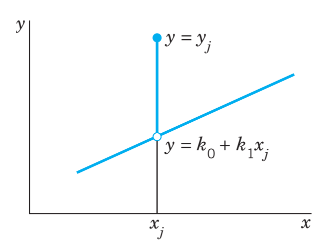


Fig. 543. Vertical distance of a point (x_j, y_j) from a straight line $y = k_0 + k_1 x$

Hence the sum of the squares of these distances is

$$q = \sum_{j=1}^{n} (y_j - k_0 - k_1 x_j)^2. \tag{3}$$

In the method of least squares we now have to determine k_0 and k_1 such that q is minimum. From calculus we know that a necessary condition for this is

$$\frac{\partial q}{\partial k_0} = 0 \qquad \text{and} \qquad \frac{\partial q}{\partial k_1} = 0. \tag{4}$$

We shall see that from this condition we obtain for the sample regression line the formula

$$y - \bar{y} = k_1 (x - \bar{x}). \tag{5}$$

Here $\bar{x}$ and $\bar{y}$ are the means of the x- and the y-values in our sample, that is,

$$(6)\qquad \text{(a)}\quad \bar{x} = \frac{1}{n}(x_1 + \cdots + x_n) \qquad\qquad \text{(b)}\quad \bar{y} = \frac{1}{n}(y_1 + \cdots + y_n).$$

The slope k_1 in (5) is called the **regression coefficient** of the sample and is given by

$$(7)\qquad k_1 = \frac{s_{xy}}{s_x^2}.$$

Here the "**sample covariance**" s_{xy} is

$$(8)\quad s_{xy} = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y}) = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j y_j - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{j=1}^{n} y_j\right)\right]$$

and s_x^2 is given by

$$(9a)\qquad s_x^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} x_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} x_j\right)^2\right].$$

From (5) we see that the sample regression line passes through the point $(\bar{x}, \bar{y})$, by which it is determined, together with the regression coefficient (7). We may call s_x^2 the *variance* of the x-values, but we should keep in mind that x is an ordinary variable, not a random variable.

We shall soon also need

$$(9b)\qquad s_y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(y_j - \bar{y})^2 = \frac{1}{n-1}\left[\sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2\right].$$

Derivation of (5) and (7). Differentiating (3) and using (4), we first obtain

$$\frac{\partial q}{\partial k_0} = -2\sum (y_j - k_0 - k_1 x_j) = 0,$$
$$\frac{\partial q}{\partial k_1} = -2\sum x_j(y_j - k_0 - k_1 x_j) = 0$$

where we sum over j from 1 to n. We now divide by 2, write each of the two sums as three sums, and take the sums containing y_j and $x_j y_j$ over to the right. Then we get the "**normal equations**"

$$(10)\qquad \begin{aligned} k_0 n + k_1 \sum x_j &= \sum y_j \\ k_0 \sum x_j + k_1 \sum x_j^2 &= \sum x_j y_j. \end{aligned}$$

This is a linear system of two equations in the two unknowns k_0 and k_1. Its coefficient determinant is [see (9)]

$$\begin{vmatrix} n & \sum x_j \\ \sum x_j & \sum x_j^2 \end{vmatrix} = n\sum x_j^2 - \left(\sum x_j\right)^2 = n(n-1)s_x^2 = n\sum (x_j - \bar{x})^2$$

and is not zero because of Assumption (A1). Hence the system has a unique solution. Dividing the first equation of (10) by n and using (6), we get $k_0 = \bar{y} - k_1\bar{x}$. Together with $y = k_0 + k_1 x$ in (2) this gives (5). To get (7), we solve the system (10) by Cramer's rule (Sec. 7.6) or elimination, finding

$$k_1 = \frac{n\sum x_j y_j - \sum x_i \sum y_j}{n(n-1)s_x^2}. \tag{11}$$

This gives (7)–(9) and completes the derivation. [The equality of the two expressions in (8) and in (9) may be shown by the student]. ■

EXAMPLE 1 **Regression Line**

The decrease of volume y [%] of leather for certain fixed values of high pressure x [atmospheres] was measured. The results are shown in the first two columns of Table 25.11. Find the regression line of y on x.

Solution. We see that $n = 4$ and obtain the values $\bar{x} = 28000/4 = 7000$, $\bar{y} = 19.0/4 = 4.75$, and from (9) and (8)

Table 25.11 Regression of the Decrease of Volume *y* [%] of Leather on the Pressure *x* [Atmospheres]

Given Values		Auxiliary Values	
x_j	y_j	x_j^2	$x_j y_j$
4000	2.3	16,000,000	9200
6000	4.1	36,000,000	24,600
8000	5.7	64,000,000	45,600
10,000	6.9	100,000,000	69,000
28,000	19.0	216,000,000	148,400

$$s_x^2 = \frac{1}{3}\left(216{,}000{,}000 - \frac{28{,}000^2}{4}\right) = \frac{20{,}000{,}000}{3}$$

$$s_{xy} = \frac{1}{3}\left(148{,}400 - \frac{28{,}000 \cdot 19}{4}\right) = \frac{15{,}400}{3}.$$

Hence $k_1 = 15{,}400/20{,}000{,}000 = 0.00077$ from (7), and the regression line is

$$y - 4.75 = 0.00077(x - 7000) \qquad \text{or} \qquad y = 0.00077x - 0.64.$$

Note that $y(0) = -0.64$, which is physically meaningless, but typically indicates that a linear relation is merely an approximation valid on some restricted interval. ■

Confidence Intervals in Regression Analysis

If we want to get confidence intervals, we have to make assumptions about the distribution of Y (which we have not made so far; least squares is a "geometric principle," nowhere involving probabilities!). We assume normality and independence in sampling:

Assumption (A2)

For each fixed x the random variable Y is normal with mean (1)*, that is,*

$$\mu(x) = \kappa_0 + \kappa_1 x \tag{12}$$

and variance σ^2 independent of x.

Assumption (A3)

The n performances of the experiment by which we obtain a sample

$$(x_1, y_1), \quad (x_2, y_2), \quad \cdots, \quad (x_n, y_n)$$

are independent.

κ_1 in (12) is called the **regression coefficient** of the population because it can be shown that, under Assumptions (A1)–(A3), the maximum likelihood estimate of κ_1 is the sample regression coefficient k_1 given by (11).

Under Assumptions (A1)–(A3), we may now obtain a confidence interval for κ_1, as shown in Table 25.12.

Table 25.12 Determination of a Confidence Interval for κ_1 in (1) under Assumptions (A1)–(A3)

Step 1. Choose a confidence level γ (95%, 99%, or the like).

Step 2. Determine the solution c of the equation

$$F(c) = \tfrac{1}{2}(1 + \gamma) \tag{13}$$

from the table of the t-distribution with $n - 2$ degrees of freedom (Table A9 in App. 5; n = sample size).

Step 3. Using a sample $(x_1, y_1), \cdots, (x_n, y_n)$, compute $(n-1)s_x^2$ from (9a), $(n-1)s_{xy}$ from (8), k_1 from (7),

$$(n-1)s_y^2 = \sum_{j=1}^{n} y_j^2 - \frac{1}{n}\left(\sum_{j=1}^{n} y_j\right)^2 \tag{14}$$

[as in (9b)], and

$$q_0 = (n-1)(s_y^2 - k_1^2 s_x^2). \tag{15}$$

Step 4. Compute

$$K = c\sqrt{\frac{q_0}{(n-2)(n-1)s_x^2}}.$$

The confidence interval is

$$\text{CONF}_\gamma \{k_1 - K \leqq \kappa_1 \leqq k_1 + K\}. \tag{16}$$

EXAMPLE 2 **Confidence Interval for the Regression Coefficient**

Using the sample in Table 25.11, determine a confidence interval for κ_1 by the method in Table 25.12.

Solution. ***Step 1.*** We choose $\gamma = 0.95$.

Step 2. Equation (13) takes the form $F(c) = 0.975$, and Table A9 in App. 5 with $n - 2 = 2$ degrees of freedom gives $c = 4.30$.

Step 3. From Example 1 we have $3s_x^2 = 20{,}000{,}000$ and $k_1 = 0.00077$. From Table 25.11 we compute

$$3s_y^2 = 102.0 - \frac{19^2}{4}$$

$$= 11.95.$$

$$q_0 = 11.95 - 20{,}000{,}000 \cdot 0.00077^2$$

$$= 0.092.$$

Step 4. We thus obtain

$$K = 4.30\sqrt{0.092/(2 \cdot 20{,}000{,}000)}$$

$$= 0.000206$$

and

$$\text{CONF}_{0.95}\{0.00056 \leqq \kappa_1 \leqq 0.00098\}.$$

Correlation Analysis

We shall now give an introduction to the basic facts in correlation analysis; for proofs see Ref. [G2] or [G8] in App. 1.

Correlation analysis is concerned with the relation between X and Y in a two-dimensional random variable (X, Y) (Sec. 24.9). A sample consists of n ordered pairs of values $(x_1, y_1), \cdots, (x_n, y_n)$, as before. The interrelation between the x and y values in the sample is measured by the sample covariance s_{xy} in (8) or by the sample **correlation coefficient**

(17)
$$r = \frac{s_{xy}}{s_x s_y}$$

with s_x and s_y given in (9). Here r has the advantage that it does not change under a multiplication of the x and y values by a factor (in going from feet to inches, etc.).

THEOREM 1 **Sample Correlation Coefficient**

The sample correlation coefficient r satisfies $-1 \leqq r \leqq 1$. In particular, $r = \pm 1$ if and only if the sample values lie on a straight line. (See Fig. 544.)

The theoretical counterpart of r is the **correlation coefficient** ρ of X and Y,

(18)
$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}$$

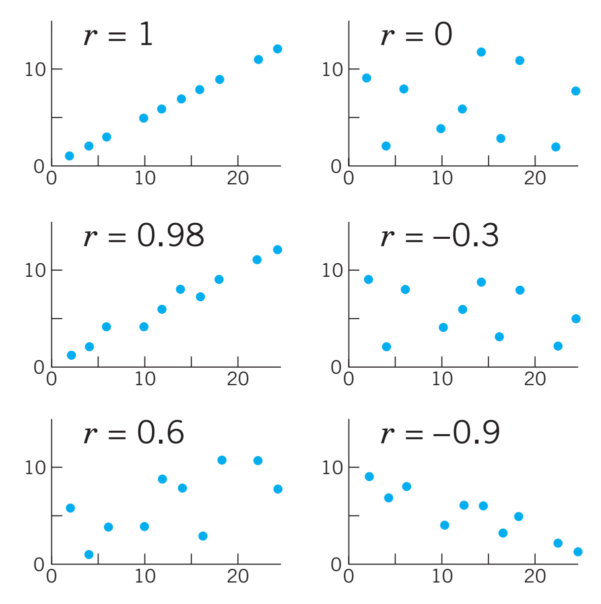


Fig. 544. Samples with various values of the correlation coefficient r

where $\mu_X = E(X)$, $\mu_Y = E(Y)$, $\sigma_X^2 = E([X - \mu_X]^2)$, $\sigma_Y^2 = E([Y - \mu_Y]^2)$ (the means and variances of the marginal distributions of X and Y; see Sec. 24.9), and σ_{XY} is the **covariance** of X and Y given by (see Sec. 24.9)

$$\sigma_{XY} = E([X - \mu_X][Y - \mu_Y]) = E(XY) - E(X)E(Y). \tag{19}$$

The analog of Theorem 1 is

THEOREM 2

Correlation Coefficient

The correlation coefficient ρ *satisfies* $-1 \leqq \rho \leqq 1$. *In particular,* $\rho = \pm 1$ *if and only if* X *and* Y *are* **linearly related**, *that is,* $Y = \gamma X + \delta$, $X = \gamma^* Y + \delta^*$.

X and Y are called **uncorrelated** if $\rho = 0$.

THEOREM 3

Independence. Normal Distribution

(a) *Independent* X *and* Y (see Sec. 24.9) *are uncorrelated.*

(b) *If* (X, Y) *is normal* (see below), *then uncorrelated* X *and* Y *are independent.*

Here the two-dimensional normal distribution can be introduced by taking two independent standardized normal random variables X^*, Y^*, whose joint distribution thus has the density

$$f^*(x^*, y^*) = \frac{1}{2\pi} e^{-(x^{*2} + y^{*2})/2} \tag{20}$$

(representing a surface of revolution over the x^*y^*-plane with a bell-shaped curve as cross section) and setting

$$X = \mu_X + \sigma_X X^*$$
$$Y = \mu_Y + \rho\sigma_Y X^* + \sqrt{1-\rho^2}\,\sigma_Y Y^*.$$

This gives the general **two-dimensional normal distribution** with the density

$$\text{(21a)} \qquad f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}} e^{-h(x,y)/2}$$

where

$$\text{(21b)} \quad h(x, y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right].$$

In Theorem 3(b), normality is important, as we can see from the following example.

EXAMPLE 3 Uncorrelated But Dependent Random Variables

If X assumes $-1, 0, 1$ with probability $\frac{1}{3}$ and $Y = X^2$, then $E(X) = 0$ and in (3)

$$\sigma_{XY} = E(XY) = E(X^3) = (-1)^3 \cdot \tfrac{1}{3} + 0^3 \cdot \tfrac{1}{3} + 1^3 \cdot \tfrac{1}{3} = 0,$$

so that $\rho = 0$ and X and Y are uncorrelated. But they are certainly not independent since they are even functionally related. ■

Test for the Correlation Coefficient ρ

Table 25.13 shows a test for ρ in the case of the two-dimensional normal distribution. t is an observed value of a random variable that has a t-distribution with $n - 2$ degrees of freedom. This was shown by R. A. Fisher (*Biometrika* **10** (1915), 507–521).

Table 25.13 Test of the Hypothesis $\rho = 0$ Against the Alternative $\rho > 0$ in the Case of the Two-Dimensional Normal Distribution

Step 1. Choose a significance level α (5%, 1%, or the like).

Step 2. Determine the solution c of the equation

$$P(T \leqq c) = 1 - \alpha$$

from the t-distribution (Table A9 in App. 5) with $n - 2$ degrees of freedom.

Step 3. Compute r from (17), using a sample $(x_1, y_1), \cdots, (x_n, y_n)$.

Step 4. Compute

$$t = r\left(\sqrt{\frac{n-2}{1-r^2}}\right).$$

If $t \leqq c$, accept the hypothesis. If $t > c$, reject the hypothesis.

EXAMPLE 4 **Test for the Correlation Coefficient ρ**

Test the hypothesis $\rho = 0$ (independence of X and Y, because of Theorem 3) against the alternative $\rho > 0$, using the data in the lower left corner of Fig. 544, where $r = 0.6$ (manual soldering errors on 10 two-sided circuit boards done by 10 workers; x = front, y = back of the boards).

Solution. We choose $\alpha = 5\%$; thus $1 - \alpha = 95\%$. Since $n = 10$, $n - 2 = 8$, the table gives $c = 1.86$. Also, $t = 0.6\sqrt{8/0.64} = 2.12 > c$. We reject the hypothesis and assert that there is a **positive correlation**. A worker making few (many) errors on the front side also tends to make few (many) errors on the reverse side of the board. ∎

PROBLEM SET 25.9

1–10 SAMPLE REGRESSION LINE

Find and graph the sample regression line of y on x and the given data as points on the same axes. Show the details of your work.

1. (0, 1.0), (2, 2.1), (4, 2.9), (6, 3.6), (8, 5.2)
2. (−2, 3.5), (1, 2.6), (3, 1.3), (5, 0.4)
3. x = Revolutions per minute, y = Power of a Diesel engine [hp]

x	400	500	600	700	750
y	5800	10,300	14,200	18,800	21,000

4. x = Deformation of a certain steel [mm], y = Brinell hardness [kg/mm^2]

x	6	9	11	13	22	26	28	33	35
y	68	67	65	53	44	40	37	34	32

5. x = Brinell hardness, y = Tensile strength [in 1000 psi (pounds per square inch)] of steel with 0.45% C tempered for 1 hour

x	200	300	400	500
y	110	150	190	280

6. **Abrasion of quenched and tempered steel S620.** x = Sliding distance [km], y = Wear volume [mm^3]

x	1.1	3.2	3.4	4.5	5.6
y	40	65	120	150	190

7. **Ohm's law (Sec. 2.9).** x = Voltage [V], y = Current [A]. Also find the resistance R [Ω].

x	40	40	80	80	110	110
y	5.1	4.8	0.0	10.3	13.0	12.7

8. **Hooke's law** (Sec. 2.4). x = Force [lb], y = Extension [in] of a spring. Also find the spring modulus.

x	2	4	6	8
y	4.1	7.8	12.3	15.8

9. **Thermal conductivity of water.** x = Temperature [°F], y = Conductivity [Btu/(hr · ft · °F)]. Also find y at room temperature 66°F.

x	32	50	100	150	212
y	0.337	0.345	0.365	0.380	0.395

10. **Stopping distance of a car.** x = Speed [mph]. y = Stopping distance [ft]. Also find y at 35 mph.

x	30	40	50	60
y	160	240	330	435

11. **CAS EXPERIMENT. Moving Data.** Take a sample, for instance, that in Prob. 4, and investigate and graph the effect of changing y-values **(a)** for small x, **(b)** for large x, **(c)** in the middle of the sample.

12–15 CONFIDENCE INTERVALS

Find a 95% confidence interval for the regression coefficient κ_1, assuming (A2) and (A3) hold and using the sample.

12. In Prob. 2
13. In Prob. 3
14. In Prob. 4
15. x = Humidity of air [%], y = Expansion of gelatin [%],

x	10	20	30	40
y	0.8	1.6	2.3	2.8

CHAPTER 25 REVIEW QUESTIONS AND PROBLEMS

1. What is a sample? A population? Why do we sample in statistics?
2. If we have several samples from the same population, do they have the same sample distribution function? The same mean and variance?
3. Can we develop statistical methods without using probability theory? Apply the methods without using a sample?
4. What is the idea of the maximum likelihood method? Why do we say "likelihood" rather than "probability"?

5. Couldn't we make the error of interval estimation zero simply by choosing the confidence level 1?

6. What is testing? Why do we test? What are the errors involved?

7. When did we use the t-distribution? The F-distribution?

8. What is the chi-square (χ^2) test? Give a sample example from memory.

9. What are one-sided and two-sided tests? Give typical examples.

10. How do we test in quality control? In acceptance sampling?

11. What is the power of a test? What could you perhaps do when it is low?

12. What is Gauss's least squares principle (which he found at age 18)?

13. What is the difference between regression and correlation?

14. Find the mean, variance, and standard derivation of the sample 21.0 21.6 19.9 19.6 15.6 20.6 22.1 22.2.

15. Assuming normality, find the maximum likelihood estimates of mean and variance from the sample in Prob. 14.

16. Determine a 95% confidence interval for the mean μ of a normal population with variance $\sigma^2 = 25$, using a sample of size 500 with mean 22.

17. Determine a 99% confidence interval for the mean of a normal population, using the sample 32, 33, 32, 34, 35, 29, 29, 27.

18. Assuming normality, find a 95% confidence interval for the variance from the sample 145.3, 145.1, 145.4, 146.2.

19. Using a sample of 10 values with mean 14.5 from a normal population with variance $\sigma^2 = 0.25$, test the hypothesis $\mu_0 = 15.0$ against the alternative $\mu_1 = 14.5$ on the 5% level. Find the power.

20. Three specimens of high-quality concrete had compressive strength 357, 359, 413 [kg/cm^2], and for three specimens of ordinary concrete the values were 346, 358, 302. Test for equality of the population means, $\mu_1 = \mu_2$, against the alternative $\mu_1 > \mu_2$. Assume normality and equality of variance. Choose $\alpha = 5\%$.

21. Assume the thickness X of washers to be normal with mean 2.75 mm and variance 0.00024 mm^2. Set up a control chart for μ and graph the means of the five samples (2.74, 2.76), (2.74, 2.74), (2.79, 2.81), (2.78, 2.76), (2.71, 2.75) on the chart.

22. The OC curve in acceptance sampling cannot have a strictly vertical portion. Why?

23. Find the risks in the sampling plan with $n = 6$ and $c = 0$, assuming that the AQL is $\theta_0 = 1\%$ and the RQL is $\theta_1 = 15\%$. How do the risks change if we increase n?

24. Does a process of producing plastic rods of length $\tilde{\mu} = 2$ meters need adjustment if in a sample, 2 rods have the exact length and 15 are shorter and 3 longer than 2 meters? (Use the sign test.)

25. Find the regression line of y on x for the data $(x, y) = (0, 4), (2, 0), (4, -5), (6, -9), (8, -10)$.

SUMMARY OF CHAPTER 25
Mathematical Statistics

We recall from Chap. 24 that, with an experiment in which we observe some quantity (number of defectives, height of persons, etc.), there is associated a random variable X whose probability distribution is given by a distribution function

$$F(x) = P(X \leqq x) \qquad (1) \quad \text{(Sec. 24.5)}$$

which for each x gives the probability that X assumes any value not exceeding x.

In statistics we take random samples $x_1, \cdots, x_n$ of size n by performing that experiment n times (Sec. 25.1) and draw conclusions from properties of samples about properties of the distribution of the corresponding X. We do this by calculating *point estimates* or *confidence intervals* or by performing a *test* for **parameters** (μ and σ^2 in the normal distribution, p in the binomial distribution, etc.) or by a test for distribution functions.

A **point estimate** (Sec. 25.2) is an approximate value for a parameter in the distribution of X obtained from a sample. Notably, the **sample mean** (Sec. 25.1)

$$\bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j = \frac{1}{n}(x_1 + \cdots + x_n) \tag{2}$$

is an estimate of the mean μ of X, and the **sample variance** (Sec. 25.1)

$$s^2 = \frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2 = \frac{1}{n-1}[(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2] \tag{3}$$

is an estimate of the variance σ^2 of X. Point estimation can be done by the basic ***maximum likelihood method*** (Sec. 25.2).

Confidence intervals (Sec. 25.3) are intervals $\theta_1 \leqq \theta \leqq \theta_2$ with endpoints calculated from a sample such that, with a high probability γ, we obtain an interval that contains the unknown true value of the parameter θ in the distribution of X. Here, γ is chosen at the beginning, usually 95% or 99%. We denote such an interval by $\text{CONF}_{\gamma}\,\{\theta_1 \leqq \theta \leqq \theta_2\}$.

In a **test** for a parameter we test a *hypothesis* $\theta = \theta_0$ against an *alternative* $\theta = \theta_1$ and then, on the basis of a sample, accept the hypothesis, or we reject it in favor of the alternative (Sec. 25.4). Like any conclusion about X from samples, this may involve errors leading to a false decision. There is a small probability α (which we can choose, 5% or 1%, for instance) that we reject a true hypothesis, and there is a probability β (which we can compute and decrease by taking larger samples) that we accept a false hypothesis. α is called the **significance level** and $1 - \beta$ the **power** of the test. Among many other engineering applications, testing is used in ***quality control*** (Sec. 25.5) and ***acceptance sampling*** (Sec. 25.6).

If not merely a parameter but the kind of distribution of X is unknown, we can use the **chi-square test** (Sec. 25.7) for testing the hypothesis that some function $F(x)$ is the unknown distribution function of X. This is done by determining the discrepancy between $F(x)$ and the distribution function $\tilde{F}(x)$ of a given sample.

"Distribution-free" or ***nonparametric tests*** are tests that apply to any distribution, since they are based on combinatorial ideas. These tests are usually very simple. Two of them are discussed in Sec. 25.8.

The last section deals with samples of ***pairs of values***, which arise in an experiment when we simultaneously observe two quantities. In ***regression analysis***, one of the quantities, x, is an ordinary variable and the other, Y, is a random variable whose mean μ depends on x, say, $\mu(x) = \kappa_0 + \kappa_1 x$. In ***correlation analysis*** the relation between X and Y in a two-dimensional random variable (X, Y) is investigated, notably in terms of the ***correlation coefficient*** ρ.

APPENDIX 1

References

Software *see* at the beginning of Chaps. 19 and 24.

General References

[GenRef1] Abramowitz, M. and I. A. Stegun (eds.), *Handbook of Mathematical Functions.* 10th printing, with corrections. Washington, DC: National Bureau of Standards. 1972 (also New York: Dover, 1965). See also [W1]

[GenRef2] Cajori, F., *History of Mathematics.* 5th ed. Reprinted. Providence, RI: American Mathematical Society, 2002.

[GenRef3] Courant, R. and D. Hilbert, *Methods of Mathematical Physics.* 2 vols. Hoboken, NJ: Wiley, 1989.

[GenRef4] Courant, R., *Differential and Integral Calculus.* 2 vols. Hoboken, NJ: Wiley, 1988.

[GenRef5] Graham, R. L. et al., *Concrete Mathematics.* 2nd ed. Reading, MA: Addison-Wesley, 1994.

[GenRef6] Ito, K. (ed.), *Encyclopedic Dictionary of Mathematics.* 4 vols. 2nd ed. Cambridge, MA: MIT Press, 1993.

[GenRef7] Kreyszig, E., *Introductory Functional Analysis with Applications.* New York: Wiley, 1989.

[GenRef8] Kreyszig, E., *Differential Geometry.* Mineola, NY: Dover, 1991.

[GenRef9] Kreyszig, E. *Introduction to Differential Geometry and Riemannian Geometry.* Toronto: University of Toronto Press, 1975.

[GenRef10] Szegö, G., *Orthogonal Polynomials.* 4th ed. Reprinted. New York: American Mathematical Society, 2003.

[GenRef11] Thomas, G. et al., *Thomas' Calculus, Early Transcendentals Update.* 10th ed. Reading, MA: Addison-Wesley, 2003.

Part A. Ordinary Differential Equations (ODEs) (Chaps. 1–6)

See also Part E: Numeric Analysis

[A1] Arnold, V. I., *Ordinary Differential Equations.* 3rd ed. New York: Springer, 2006.

[A2] Bhatia, N. P. and G. P. Szego, *Stability Theory of Dynamical Systems.* New York: Springer, 2002.

[A3] Birkhoff, G. and G.-C. Rota, *Ordinary Differential Equations.* 4th ed. New York: Wiley, 1989.

[A4] Brauer, F. and J. A. Nohel, *Qualitative Theory of Ordinary Differential Equations.* Mineola, NY: Dover, 1994.

[A5] Churchill, R. V., *Operational Mathematics.* 3rd ed. New York: McGraw-Hill, 1972.

[A6] Coddington, E. A. and R. Carlson, *Linear Ordinary Differential Equations.* Philadelphia: SIAM, 1997.

[A7] Coddington, E. A. and N. Levinson, *Theory of Ordinary Differential Equations.* Malabar, FL: Krieger, 1984.

[A8] Dong, T.-R. et al., *Qualitative Theory of Differential Equations.* Providence, RI: American Mathematical Society, 1992.

[A9] Erdélyi, A. et al., *Tables of Integral Transforms.* 2 vols. New York: McGraw-Hill, 1954.

[A10] Hartman, P., *Ordinary Differential Equations.* 2nd ed. Philadelphia: SIAM, 2002.

[A11] Ince, E. L., *Ordinary Differential Equations.* New York: Dover, 1956.

[A12] Schiff, J. L., *The Laplace Transform: Theory and Applications.* New York: Springer, 1999.

[A13] Watson, G. N., *A Treatise on the Theory of Bessel Functions.* 2nd ed. Reprinted. New York: Cambridge University Press, 1995.

[A14] Widder, D. V., *The Laplace Transform.* Princeton, NJ: Princeton University Press, 1941.

[A15] Zwillinger, D., *Handbook of Differential Equations.* 3rd ed. New York: Academic Press, 1998.

Part B. Linear Algebra, Vector Calculus (Chaps. 7–10)

For books on *numeric* linear algebra, *see also* Part E: Numeric Analysis.

[B1] Bellman, R., *Introduction to Matrix Analysis.* 2nd ed. Philadelphia: SIAM, 1997.

[B2] Chatelin, F., *Eigenvalues of Matrices.* New York: Wiley-Interscience, 1993.

[B3] Gantmacher, F. R., *The Theory of Matrices.* 2 vols. Providence, RI: American Mathematical Society, 2000.

[B4] Gohberg, I. P. et al., *Invariant Subspaces of Matrices with Applications.* New York: Wiley, 2006.

[B5] Greub, W. H., *Linear Algebra.* 4th ed. New York: Springer, 1975.

[B6] Herstein, I. N., *Abstract Algebra.* 3rd ed. New York: Wiley, 1996.

[B7] Joshi, A. W., *Matrices and Tensors in Physics.* 3rd ed. New York: Wiley, 1995.
[B8] Lang, S., *Linear Algebra.* 3rd ed. New York: Springer, 1996.
[B9] Nef, W., *Linear Algebra.* 2nd ed. New York: Dover, 1988.
[B10] Parlett, B., *The Symmetric Eigenvalue Problem.* Philadelphia: SIAM, 1998.

Part C. Fourier Analysis and PDEs (Chaps. 11–12)

For books on *numerics* for PDEs *see also* Part E: Numeric Analysis.

[C1] Antimirov, M. Ya., *Applied Integral Transforms.* Providence, RI: American Mathematical Society, 1993.
[C2] Bracewell, R., *The Fourier Transform and Its Applications.* 3rd ed. New York: McGraw-Hill, 2000.
[C3] Carslaw, H. S. and J. C. Jaeger, *Conduction of Heat in Solids.* 2nd ed. Reprinted. Oxford: Clarendon, 2000.
[C4] Churchill, R. V. and J. W. Brown, *Fourier Series and Boundary Value Problems.* 6th ed. New York: McGraw-Hill, 2006.
[C5] DuChateau, P. and D. Zachmann, *Applied Partial Differential Equations.* Mineola, NY: Dover, 2002.
[C6] Hanna, J. R. and J. H. Rowland, *Fourier Series, Transforms, and Boundary Value Problems.* 2nd ed. New York: Wiley, 2008.
[C7] Jerri, A. J., *The Gibbs Phenomenon in Fourier Analysis, Splines, and Wavelet Approximations.* Boston: Kluwer, 1998.
[C8] John, F., *Partial Differential Equations.* 4th edition New York: Springer, 1982.
[C9] Tolstov, G. P., *Fourier Series.* New York: Dover, 1976.
[C10] Widder, D. V., *The Heat Equation.* New York: Academic Press, 1975.
[C11] Zauderer, E., *Partial Differential Equations of Applied Mathematics.* 3rd ed. New York: Wiley, 2006.
[C12] Zygmund, A. and R. Fefferman, *Trigonometric Series.* 3rd ed. New York: Cambridge University Press, 2002.

Part D. Complex Analysis (Chaps. 13–18)

[D1] Ahlfors, L. V., *Complex Analysis.* 3rd ed. New York: McGraw-Hill, 1979.
[D2] Bieberbach, L., *Conformal Mapping.* Providence, RI: American Mathematical Society, 2000.
[D3] Henrici, P., *Applied and Computational Complex Analysis.* 3 vols. New York: Wiley, 1993.
[D4] Hille, E., *Analytic Function Theory.* 2 vols. 2nd ed. Providence, RI: American Mathematical Society, Reprint V1 1983, V2 2005.
[D5] Knopp, K., *Elements of the Theory of Functions.* New York: Dover, 1952.
[D6] Knopp, K., *Theory of Functions.* 2 parts. New York: Dover, Reprinted 1996.
[D7] Krantz, S. G., *Complex Analysis: The Geometric Viewpoint.* Washington, DC: The Mathematical Association of America, 1990.
[D8] Lang, S., *Complex Analysis.* 4th ed. New York: Springer, 1999.
[D9] Narasimhan, R., *Compact Riemann Surfaces.* New York: Springer, 1996.
[D10] Nehari, Z., *Conformal Mapping.* Mineola, NY: Dover, 1975.
[D11] Springer, G., *Introduction to Riemann Surfaces.* Providence, RI: American Mathematical Society, 2001.

Part E. Numeric Analysis (Chaps. 19–21)

[E1] Ames, W. F., *Numerical Methods for Partial Differential Equations.* 3rd ed. New York: Academic Press, 1992.
[E2] Anderson, E., et al., *LAPACK User's Guide.* 3rd ed. Philadelphia: SIAM, 1999.
[E3] Bank, R. E., *PLTMG. A Software Package for Solving Elliptic Partial Differential Equations: Users' Guide 8.0.* Philadelphia: SIAM, 1998.
[E4] Constanda, C., *Solution Techniques for Elementary Partial Differential Equations.* Boca Raton, FL: CRC Press, 2002.
[E5] Dahlquist, G. and A. Björck, *Numerical Methods.* Mineola, NY: Dover, 2003.
[E6] DeBoor, C., *A Practical Guide to Splines.* Reprinted. New York: Springer, 2001.
[E7] Dongarra, J. J. et al., *LINPACK Users Guide.* Philadelphia: SIAM, 1979. (See also at the beginning of Chap. 19.)
[E8] Garbow, B. S. et al., *Matrix Eigensystem Routines: EISPACK Guide Extension.* Reprinted. New York: Springer, 1990.
[E9] Golub, G. H. and C. F. Van Loan, *Matrix Computations.* 3rd ed. Baltimore, MD: Johns Hopkins University Press, 1996.
[E10] Higham, N. J., *Accuracy and Stability of Numerical Algorithms.* 2nd ed. Philadelphia: SIAM, 2002.
[E11] IMSL (International Mathematical and Statistical Libraries), *FORTRAN Numerical Library.* Houston, TX: Visual Numerics, 2002. (See also at the beginning of Chap. 19.)
[E12] IMSL, *IMSL for Java.* Houston, TX: Visual Numerics, 2002.
[E13] IMSL, *C Library.* Houston, TX: Visual Numerics, 2002.
[E14] Kelley, C. T., *Iterative Methods for Linear and Nonlinear Equations.* Philadelphia: SIAM, 1995.
[E15] Knabner, P. and L. Angerman, *Numerical Methods for Partial Differential Equations.* New York: Springer, 2003.

[E16] Knuth, D. E., *The Art of Computer Programming*. 3 vols. 3rd ed. Reading, MA: Addison-Wesley, 1997–2009.

[E17] Kreyszig, E., *Introductory Functional Analysis with Applications*. New York: Wiley, 1989.

[E18] Kreyszig, E., On methods of Fourier analysis in multigrid theory. *Lecture Notes in Pure and Applied Mathematics* 157. New York: Dekker, 1994, pp. 225–242.

[E19] Kreyszig, E., Basic ideas in modern numerical analysis and their origins. *Proceedings of the Annual Conference of the Canadian Society for the History and Philosophy of Mathematics*. 1997, pp. 34–45.

[E20] Kreyszig, E., and J. Todd, *QR* in two dimensions. *Elemente der Mathematik* 31 (1976), pp. 109–114.

[E21] Mortensen, M. E., *Geometric Modeling*. 2nd ed. New York: Wiley, 1997.

[E22] Morton, K. W., and D. F. Mayers, *Numerical Solution of Partial Differential Equations: An Introduction*. New York: Cambridge University Press, 1994.

[E23] Ortega, J. M., *Introduction to Parallel and Vector Solution of Linear Systems*. New York: Plenum Press, 1988.

[E24] Overton, M. L., *Numerical Computing with IEEE Floating Point Arithmetic*. Philadelphia: SIAM, 2004.

[E25] Press, W. H. et al., *Numerical Recipes in C: The Art of Scientific Computing*. 2nd ed. New York: Cambridge University Press, 1992.

[E26] Shampine, L. F., *Numerical Solutions of Ordinary Differential Equations*. New York: Chapman and Hall, 1994.

[E27] Varga, R. S., *Matrix Iterative Analysis*. 2nd ed. New York: Springer, 2000.

[E28] Varga, R. S., *Geršgorin and His Circles*. New York: Springer, 2004.

[E29] Wilkinson, J. H., *The Algebraic Eigenvalue Problem*. Oxford: Oxford University Press, 1988.

Part F. Optimization, Graphs (Chaps. 22–23)

[F1] Bondy, J. A. and U.S.R. Murty, *Graph Theory with Applications*. Hoboken, NJ: Wiley-Interscience, 1991.

[F2] Cook, W. J. et al., *Combinatorial Optimization*. New York: Wiley, 1997.

[F3] Diestel, R., *Graph Theory*. 4th ed. New York: Springer, 2006.

[F4] Diwekar, U. M., *Introduction to Applied Optimization*. 2nd ed. New York: Springer, 2008.

[F5] Gass, S. L., *Linear Programming. Method and Applications*. 3rd ed. New York: McGraw-Hill, 1969.

[F6] Gross, J. T. and J.Yellen (eds.), *Handbook of Graph Theory and Applications*. 2nd ed. Boca Raton, FL: CRC Press, 2006.

[F7] Goodrich, M. T., and R. Tamassia, *Algorithm Design: Foundations, Analysis, and Internet Examples*. Hoboken, NJ: Wiley, 2002.

[F8] Harary, F., *Graph Theory*. Reprinted. Reading, MA: Addison-Wesley, 2000.

[F9] Merris, R., *Graph Theory*. Hoboken, NJ: Wiley-Interscience, 2000.

[F10] Ralston, A., and P. Rabinowitz, *A First Course in Numerical Analysis*. 2nd ed. Mineola, NY: Dover, 2001.

[F11] Thulasiraman, K., and M. N. S. Swamy, *Graph Theory and Algorithms*. New York: Wiley-Interscience, 1992.

[F12] Tucker, A., *Applied Combinatorics*. 5th ed. Hoboken, NJ: Wiley, 2007.

Part G. Probability and Statistics (Chaps. 24–25)

[G1] American Society for Testing Materials, *Manual on Presentation of Data and Control Chart Analysis*. 7th ed. Philadelphia: ASTM, 2002.

[G2] Anderson, T. W., *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Hoboken, NJ: Wiley, 2003.

[G3] Cramér, H., *Mathematical Methods of Statistics*. Reprinted. Princeton, NJ: Princeton University Press, 1999.

[G4] Dodge, Y., *The Oxford Dictionary of Statistical Terms*. 6th ed. Oxford: Oxford University Press, 2006.

[G5] Gibbons, J. D. and S. Chakraborti, *Nonparametric Statistical Inference*. 4th ed. New York: Dekker, 2003.

[G6] Grant, E. L. and R. S. Leavenworth, *Statistical Quality Control*. 7th ed. New York: McGraw-Hill, 1996.

[G7] IMSL, *Fortran Numerical Library*. Houston, TX: Visual Numerics, 2002.

[G8] Kreyszig, E., *Introductory Mathematical Statistics. Principles and Methods*. New York: Wiley, 1970.

[G9] O'Hagan, T. et al., *Kendall's Advanced Theory of Statistics 3-Volume Set*. Kent, U.K.: Hodder Arnold, 2004.

[G10] Rohatgi, V. K. and A. K. MD. E. Saleh, *An Introduction to Probability and Statistics*. 2nd ed. Hoboken, NJ: Wiley-Interscience, 2001.

Web References

[W1] upgraded version of [GenRef1] online at http://dlmf.nist.gov/. Hardcopy and CD-Rom: Oliver, W. J. et al. (eds.), *NIST Handbook of Mathematical Functions*. Cambridge; New York: Cambridge University Press, 2010.

[W2] O'Connor, J. and E. Robertson, MacTutor History of Mathematics Archive. St. Andrews, Scotland: University of St. Andrews, School of Mathematics and Statistics. Online at http://www-history.mcs.st-andrews.ac.uk. (Biographies of mathematicians, etc.).

APPENDIX 2

Answers to Odd-Numbered Problems

Problem Set 1.1, page 8

1. $y = \dfrac{1}{\pi}\cos 2\pi x + c$

3. $y = ce^{x}$

5. $y = 2e^{-x}(\sin x - \cos x) + c$

7. $y = \dfrac{1}{5.13}\sinh 5.13x + c$

9. $y = 1.65e^{-4x} + 0.35$

11. $y = (x + \frac{1}{2})e^{x}$

13. $y = 1/(1 + 3e^{-x})$

15. $y = 0$ and $y = 1$ because $y' = 0$ for these y

17. $\exp(-1.4 \cdot 10^{-11}t) = \frac{1}{2}$, $t = 10^{11}(\ln 2)/1.4$ [sec]

19. Integrate $y'' = g$ twice, $y'(t) = gt + v_0$, $y'(0) = v_0 = 0$ (start from rest), then $y(t) = \frac{1}{2}gt^2 + y_0$, where $y(0) = y_0 = 0$

Problem Set 1.2, page 11

11. Straight lines parallel to the x-axis

13. $y = x$

15. $mv' = mg - bv^2$, $v' = 9.8 - v^2$, $v(0) = 10$, $v' = 0$ gives the limit$/9.8 = 3.1$ [meter/sec]

17. Errors of steps 1, 5, 10: 0.0052, 0.0382, 0.1245, approximately

19. $x_5 = 0.0286$ (error 0.0093), $x_{10} = 0.2196$ (error 0.0189)

Problem Set 1.3, page 18

1. If you add a constant later, you may not get a solution.
Example: $y' = y$, $\ln|y| = x + c$, $y = e^{x+c} = \tilde{c}e^{x}$ but not $e^{x} + c$ (with $c \neq 0$)

3. $\cos^2 y\, dy = dx$, $\frac{1}{2}y + \frac{1}{4}\sin 2y + c = x$

5. $y^2 + 36x^2 = c$, ellipses

7. $y = x\arctan(x^2 + c)$

9. $y = x/(c - x)$

11. $y = 24/x$, hyperbola

13. $dy/\sin^2 y = dx/\cosh^2 x$, $-\cot y = \tanh x + c$, $c = 0$, $y = -\text{arccot}(\tanh x)$

15. $y^2 + 4x^2 = c = 25$

17. $y = x\arctan(x^3 - 1)$

19. $y_0 e^{kt} = 2y_0$, $e^{k} = 2$ (1 week), $e^{2k} = 2^2$ (2 weeks), $e^{4k} = 2^4$

21. 69.6% of y_0

23. $PV = c = \text{const}$

25. $T = 22 - 17e^{-0.5306t} = 21.9$ [°C] when $t = 9.68$ min

27. $e^{-k\cdot 10} = \frac{1}{2}$, $k = \frac{1}{10}$, $\ln\frac{1}{2}$, $e^{-kt_0} = 0.01$, $t = (\ln 100)/k = 66$ [min]

29. No. Use Newton's law of cooling.

31. $y = ax$, $y' = g(y/x) = a = \text{const}$, independent of the point (x, y)

33. $\Delta S = 0.15S\Delta\phi$, $dS/d\phi = 0.15S$, $S = S_0e^{0.15\phi} = 1000S_0$, $\phi = (1/0.15)\ln 1000 = 7.3 \cdot 2\pi$. Eight times.

Problem Set 1.4, page 26

1. Exact, $2x = 2x$, $x^2y = c$, $y = c/x^2$
3. Exact, $y = \arccos(c/\cos x)$
5. Not exact, $y = \sqrt{x^2 + cx}$
7. $F = e^{x^2}$, $e^{x^2}\tan y = c$
9. Exact, $u = e^{2x}\cos y + k(y)$, $u_y = -e^{2x}\sin y + k'$, $k' = 0$. *Ans.* $e^{2x}\cos y = 1$
11. $F = \sinh x$, $\sinh^2 x \cos y = c$
13. $u = e^x + k(y)$, $u_y = k' = -1 + e^y$, $k = -y + e^y$. *Ans.* $e^x - y + e^y = c$
15. $b = k$, $ax^2 + 2kxy + ly^2 = c$

Problem Set 1.5, page 34

3. $y = ce^x - 5.2$
5. $y = (x + c)e^{-kx}$
7. $y = x^2(c + e^x)$
9. $y = (x - 2.5/e)e^{\cos x}$
11. $y = 2 + c \sin x$
13. Separate. $y - 2.5 = c\cosh^4 1.5x$
15. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = 0 + 0 = 0$
17. $(y_1 + y_2)' + p(y_1 + y_2) = (y_1' + py_1) + (y_2' + py_2) = r + 0 = r$
19. Solution of $cy_1' + pcy_1 = c(y_1' + py_1) = cr$
21. $y = uy^*$, $y' + py = u'y^* + uy^{*\prime} + puy^* = u'y^* + u(y^{*\prime} + py^*) = u'y^* + u \cdot 0 = r$, $u' = r/y^* = re^{\int p\,dx}$, $u = \int e^{\int p\,dx} r\,dx + c$. Thus, $y = uy_h$ gives (4). We shall see that this method extends to higher-order ODEs (Secs. 2.10 and 3.3).
23. $y^2 = 1 + 8e^{-x^2}$
25. $y = 1/u$, $u = ce^{-3.2x} + 10/3.2$
27. $dx/dy = 6e^y - 2x$, $x = ce^{-2y} + 2e^y$
31. $T = 240e^{kt} + 60$, $T(10) = 200$, $k = -0.0539$, $t = 102$ min
33. $y' = A - ky$, $y(0) = 0$, $y = A(1 - e^{-kt})/k$
35. $y' = 175(0.0001 - y/450)$, $y(0) = 450 \cdot 0.0004 = 0.18$,
$y = 0.135e^{-0.3889t} + 0.045 = 0.18/2$,
$e^{-0.3889t} = (0.09 - 0.045)/0.135 = 1/3$,
$t = (\ln 3)/0.3889 = 2.82$. *Ans.* About 3 years
37. $y' = y - y^2 - 0.2y$, $y = 1/(1.25 - 0.75e^{-0.8t})$, limit 0.8, limit 1
39. $y' = By^2 - Ay = By(y - A/B)$, $A > 0$, $B > 0$. Constant solutions $y = 0$, $y = A/B$, $y' > 0$ if $y > A/B$ (unlimited growth), $y' < 0$ if $0 < y < A/B$ (extinction). $y = A/(ce^{At} + B)$, $y(0) > A/B$ if $c < 0$, $y(0) < A/B$ if $c > 0$.

Problem Set 1.6, page 38

1. $x^2/(c^2 + 9) + y^2/c^2 - 1 = 0$
3. $y - \cosh(x - c) - c = 0$
5. $y/x = c$, $y'/x = y/x^2$, $y' = y/x$, $\tilde{y}' = -x/\tilde{y}$, $\tilde{y}^2 + x^2 = \tilde{c}$, circles
7. $2\tilde{y}^2 - x^2 = \tilde{c}$
9. $y' = -2xy$, $\tilde{y}' = 1/(2x\tilde{y})$, $x = \tilde{c}e^{\tilde{y}^2}$
11. $\tilde{y} = \tilde{c}x$
13. $y' = -4x/9y$. Trajectories $\tilde{y}' = 9\tilde{y}/4x$, $\tilde{y} = \tilde{c}x^{9/4}$ ($\tilde{c} > 0$). Sketch or graph these curves.
15. $u = c$, $u_x\,dx + u_y\,dy = 0$, $y' = -u_x/u_y$. Trajectories $\tilde{y}' = u_{\tilde{y}}/u_x$. Now $v = \tilde{c}$, $v_x\,dx + v_y\,dy = 0$, $y' = -v_x/v_y$. This agrees with the trajectory ODE in u if $u_x = v_y$ (equal denominators) and $u_y = -v_x$ (equal numerators). But these are just the Cauchy–Riemann equations.

Problem Set 1.7, page 42

1. $y' = f(x, y) = r(x) - p(x)y$; hence $\partial f/\partial y = -p(x)$ is continuous and is thus bounded in the closed interval $|x - x_0| \leqq a$.

3. In $|x - x_0| < a$; just take b in $\alpha = b/K$ large, namely, $b = \alpha K$.

5. R has sides $2a$ and $2b$ and center $(1, 1)$ since $y(1) = 1$. In R, $f = 2y^2 \leqq 2(b+1)^2 = K$, $\alpha = b/K = b/(2(b+1)^2)$, $d\alpha/db = 0$ gives $b = 1$, and $\alpha_{opt} = b/K = \frac{1}{8}$. Solution by $dy/y^2 = 2\,dx$, etc., $y = 1/(3 - 2x)$.

7. $|1 + y^2| \leqq K = 1 + b^2$, $\alpha = b/K$, $d\alpha/db = 0$, $b = 1$, $\alpha = \frac{1}{2}$.

9. No. At a common point (x_1, y_1) they would both satisfy the "initial condition" $y(x_1) = y_1$, violating uniqueness.

Chapter 1 Review Questions and Problems, page 43

11. $y = ce^{-2x}$

13. $y = 1/(ce^{-4x} + 4)$

15. $y = ce^{-x} + 0.01 \cos 10x + 0.1 \sin 10x$

17. $y = ce^{-2.5x} + 0.640x - 0.256$

19. $25y^2 - 4x^2 = c$

21. $F = x$, $x^3e^y + x^2y = c$

23. $y = \sin(x + \frac{1}{4}\pi)$

25. $3 \sin x + \frac{1}{3} \sin y = 0$

27. $e^k = 1.25$, $(\ln 2)/\ln 1.25 = 3.1$, $(\ln 3)/\ln 1.25 = 4.9$ [days]

29. $e^k = 0.9$, 6.6 days. 43.7 days from $e^{kt} = 0.5$, $e^{kt} = 0.01$

Problem Set 2.1, page 53

1. $F(x, z, z') = 0$

3. $y = c_1e^{-x} + c_2$

5. $y = (c_1x + c_2)^{-1/2}$

7. $(dz/dy)z = -z^3 \sin y$, $-1/z = -dx/dy = \cos y + \tilde{c}_1$, $x = -\sin y + c_1y + c_2$

9. $y_2 = x^3 \ln x$

11. $y = c_1e^{2x} + c_2$

13. $y(t) = c_1e^{-t} + kt + c_2$

15. $y = 3 \cos 2.5x - \sin 2.5x$

17. $y = -0.75x^{3/2} - 2.25x^{-1/2}$

19. $y = 15e^{-x} - \sin x$

Problem Set 2.2, page 59

1. $y = c_1e^{-2.5x} + c_2e^{2.5x}$

3. $y = c_1e^{-2.8x} + c_2e^{-3.2x}$

5. $y = (c_1 + c_2x)e^{-\pi x}$

7. $y = c_1 + c_2e^{-4.5x}$

9. $y = c_1e^{-2.6x} + c_2e^{0.8x}$

11. $y = c_1e^{-x/2} + c_2e^{3x/2}$

13. $y = (c_1 + c_2x)e^{5x/3}$

15. $y = e^{-0.27x}(A \cos(\sqrt{\pi}\,x) + B \sin(\sqrt{\pi}\,x))$

17. $y'' + 2\sqrt{5}y' + 5y = 0$

19. $y'' + 4y' + 5y = 0$

21. $y = 4.6 \cos 5x - 0.24 \sin 5x$

23. $y = 6e^{2x} + 4e^{-3x}$

25. $y = 2e^{-x}$

27. $y = (4.5 - x)e^{-\pi x}$

29. $y = \dfrac{1}{\sqrt{\pi}} e^{-0.27x} \sin(\sqrt{\pi}\,x)$

31. Independent

33. $c_1x^2 + c_2x^2 \ln x = 0$ with $x = 1$ gives $c_1 = 0$; then $c_2 = 0$ for $x = 2$, say. Hence independent

35. Dependent since $\sin 2x = 2 \sin x \cos x$

37. $y_1 = e^{-x}$, $y_2 = 0.001e^{x} + e^{-x}$

Problem Set 2.3, page 61

1. $4e^{2x}$, $-e^{-x} + 8e^{2x}$, $-\cos x - 2 \sin x$
3. 0, 0, $(D - 2I)(-4e^{-2x}) = 8e^{-2x} + 8e^{-2x}$
5. 0, $5e^{2x}$, 0
7. $(2D - I)(2D + I)$, $y = c_1 e^{0.5x} + c_2 e^{-0.5x}$
9. $(D - 2.1I)^2$, $y = (c_1 + c_2 x)e^{2.1x}$
11. $(D - 1.6I)(D - 2.4I)$, $y = c_1 e^{1.6x} + c_2 e^{2.4x}$
15. Combine the two conditions to get $L(cy + kw) = L(cy) + L(kw) = cLy + kLw$. The converse is simple.

Problem Set 2.4, page 69

1. $y' = y_0 \cos \omega_0 t + (v_0/\omega_0) \sin \omega_0 t$. At integer t (if $\omega_0 = \pi$), because of periodicity.
3. (i) Lower by a factor $\sqrt{2}$, (ii) higher by $\sqrt{2}$
5. 0.3183, 0.4775, $\sqrt{(k_1 + k_2)/m}/(2\pi) = 0.5738$
7. $mL\theta'' = -mg \sin \theta \approx -mg\theta$ (tangential component of $W = mg$), $\theta'' + \omega_0{}^2\theta = 0$, $\omega_0/(2\pi) = \sqrt{g/L}/(2\pi)$
9. $my'' = -\tilde{a}\gamma y$, where $m = 1$ kg, $ay = \pi \cdot 0.01^2 \cdot 2y$ meter^3 is the volume of the water that causes the restoring force $a\gamma y$ with $\gamma = 9800$ nt (= weight/meter^3). $y'' + \omega_0{}^2 y = 0$, $\omega_0{}^2 = a\gamma/m = a\gamma = 0.000628\gamma$. Frequency $\omega_0/2\pi = 0.4$ [sec^{-1}].
13. $y = [y_0 + (v_0 + \alpha y_0)t]e^{-\alpha t}$, $y = [1 + (v_0 + 1)t]e^{-t}$; (ii) $v_0 = -2, -\frac{3}{2}, -\frac{4}{3}, -\frac{5}{4}, -\frac{6}{5}$
15. $\omega^* = [\omega_0{}^2 - c^2/(4m^2)]^{1/2} = \omega_0[1 - c^2/(4mk)]^{1/2} \approx \omega_0(1 - c^2/8mk) = 2.9583$
17. The positive solutions of $\tan t = 1$, that is, $\pi/4$ (max), $5\pi/4$ (min). etc
19. $0.0231 = (\ln 2)/30$ [kg/sec] from $\exp(-10 \cdot 3c/2m) = \frac{1}{2}$.

Problem Set 2.5, page 73

3. $y = (c_1 + c_2 \ln x)x^{-1.8}$
5. $\sqrt{x}\,(c_1 \cos(\ln x) + c_2 \sin(\ln x))$
7. $y = c_1 x^2 + c_2 x^3$
9. $y = (c_1 + c_2 \ln x)x^{0.6}$
11. $y = x^2(c_1 \cos(\sqrt{6} \ln x) + c_2 \sin(\sqrt{6} \ln x))$
13. $y = x^{-3/2}$
15. $y = (3.6 + 4.0 \ln x)/x$
17. $y = \cos(\ln x) + \sin(\ln x)$
19. $y = -0.525x^5 + 0.625x^{-3}$

Problem Set 2.6, page 79

3. $W = -2.2e^{-3x}$ **5.** $W = -x^4$ **7.** $W = a$
9. $y'' + 25y = 0$, $W = 5$, $y = 3 \cos 5x - \sin 5x$
11. $y'' + 5y + 6.34 = 0$, $W = 0.3e^{-5x}$, $3e^{-2.5} \cos 0.3x$
13. $y'' + 2y' = 0$, $W = -2e^{-2x}$, $y = 0.5(1 + e^{-2x})$
15. $y'' - 3.24y = 0$, $W = 1.8$, $y = 14.2 \cosh 1.8x + 9.1 \sinh 1.8x$

Problem Set 2.7, page 84

1. $y = c_1 e^{-x} + c_2 e^{-4x} - 5e^{-3x}$
3. $y = c_1 e^{-2x} + c_2 e^{-x} + 6x^2 - 18x + 21$
5. $y = (c_1 + c_2 x)e^{-2x} + \frac{1}{2}e^{-x} \sin x$
7. $y = c_1 e^{-x/2} + c_2 e^{-3x/2} + \frac{4}{5}e^x + 6x - 16$
9. $y = c_1 e^{4x} + c_2 e^{-4x} + 1.2xe^{4x} - 2e^x$
11. $y = \cos(\sqrt{3}x) + 6x^2 - 4$

13. $y = e^{x/4} - 2e^{x/2} + \frac{1}{5}e^{-x} + e^{x}$ **15.** $y = \ln x$
17. $y = e^{-0.1x}(1.5 \cos 0.5x - \sin 0.5x) + 2e^{0.5x}$

Problem Set 2.8, page 91

3. $y_p = 1.0625 \cos 2t + 3.1875 \sin 2t$
5. $y_p = -1.28 \cos 4.5t + 0.36 \sin 4.5t$
7. $y_p = 25 + \frac{4}{3} \cos 3t + \sin 3t$
9. $y = e^{-1.5t}(A \cos t + B \sin t) + 0.8 \cos t + 0.4 \sin t$
11. $y = A \cos \sqrt{2}t + B \sin \sqrt{2}t + t(\sin \sqrt{2}t - \cos \sqrt{2}t)/(2\sqrt{2})$
13. $y = A \cos t + B \sin t - (\cos \omega t)/(\omega^2 - 1)$
15. $y = e^{-2t}(A \cos 2t + B \sin 2t) + \frac{1}{4} \sin 2t$
17. $y = \frac{1}{3} \sin t - \frac{1}{15} \sin 3t - \frac{1}{105} \sin 5t$
19. $y = e^{-t}(0.4 \cos t + 0.8 \sin t) + e^{-t/2}(-0.4 \cos \frac{1}{2}t + 0.8 \sin \frac{1}{2}t)$
25. CAS Experiment. The choice of ω needs experimentation, inspection of the curves obtained, and then changes on a trail-and-error basis. It is interesting to see how in the case of beats the period gets increasingly longer and the maximum amplitude gets increasingly larger as $\omega/(2\pi)$ approaches the resonance frequency.

Problem Set 2.9, page 98

1. $RI' + I/C = 0, \quad I = ce^{-t/(RC)}$
3. $LI' + RI = E, \quad I = (E/R) + ce^{-Rt/L} = 4.8 + ce^{-40t}$
5. $I = 2(\cos t - \cos 20t)/399$
7. I_0 is maximum when $S = 0$; thus, $C = 1/(\omega^2 L)$.
9. $I = 0$ **11.** $I = 5.5 \cos 10t + 16.5 \sin 10t$ A
13. $I = e^{-5t}(A \cos 10t + B \sin 10t) - 400 \cos 25t + 200 \sin 25t$ A
15. $R > R_{\text{crit}} = 2\sqrt{L/C}$ is Case I, etc.
17. $E(0) = 600$, $I'(0) = 600$, $I = e^{-3t}(-100 \cos 4t + 75 \sin 4t) + 100 \cos t$
19. $R = 2\ \Omega, \quad L = 1$ H, $\quad C = \frac{1}{12}$ F, $\quad E = 4.4 \sin 10t$ V

Problem Set 2.10, page 102

1. $y = A \cos 3x + B \sin 3x + \frac{1}{9}(\cos 3x) \ln |\cos 3x| + \frac{1}{3}x \sin 3x$
3. $y = c_1 x + c_2 x^2 - x \sin x$ **5.** $y = A \cos x + B \sin x + \frac{1}{2}x(\cos x + \sin x)$
7. $y = (c_1 + c_2 x)e^{2x} + x^{-2}e^{2x}$ **9.** $y = (c_1 + c_2 x)e^{x} + 4x^{7/2}e^{x}$
11. $y = c_1 x^2 + c_2 x^3 + 1/(2x^4)$ **13.** $y = c_1 x^{-3} + c_2 x^3 + 3x^5$

Chapter 2 Review Questions and Problems, page 102

7. $y = c_1 e^{-4.5x} + c_2 e^{-3.5x}$ **9.** $y = e^{-3x}(A \cos 5x + B \sin 5x)$
11. $y = (c_1 + c_2 x)e^{0.8x}$ **13.** $y = c_1 x^{-4} + c_2 x^3$
15. $y = c_1 e^{2x} + c_2 e^{-x/2} - 3x + x^2$ **17.** $y = (c_1 + c_2 x)e^{1.5x} + 0.25x^2 e^{1.5x}$
19. $y = 5 \cos 4x - \frac{3}{4}\sin 4x + e^{x}$ **21.** $y = -4x + 2x^3 + 1/x$
23. $I = -0.01093 \cos 415t + 0.05273 \sin 415t$ A

25. $I = \frac{1}{73}(50 \sin 4t - 110 \cos 4t)$ A
27. *RLC*-circuit with $R = 20\ \Omega$, $L = 4$ H, $C = 0.1$ F, $E = -25 \cos 4t$ V
29. $\omega = 3.1$ is close to $\omega_0 = \sqrt{k/m} = 3$, $y = 25(\cos 3t - \cos 3.1t)$.

Problem Set 3.1, page 111

9. Linearly independent
11. Linearly independent
13. Linearly independent
15. Linearly dependent

Problem Set 3.2, page 116

1. $y = c_1 + c_2 \cos 5x + c_3 \sin 5x$
3. $y = c_1 + c_2 x + c_3 \cos 2x + c_4 \sin 2x$
5. $y = A_1 \cos x + B_1 \sin x + A_2 \cos 3x + B_2 \sin 3x$
7. $y = 2.398 + e^{-1.6x}(1.002 \cos 1.5x - 1.998 \sin 1.5x)$
9. $y = 4e^{-x} + 5e^{-x/2} \cos 3x$
11. $y = \cosh 5x - \cos 4x$
13. $y = e^{0.25x} + 4.3e^{-0.7x} + 12.1 \cos 0.1x - 0.6 \sin 0.1x$

Problem Set 3.3, page 122

1. $y = (c_1 + c_2 x + c_3 x^2)e^{-x} + \frac{1}{8}e^x - x + 2$
3. $y = c_1 \cos x + c_2 \sin x + c_3 \cos 3x + c_4 \sin 3x + 0.1 \sinh 2x$
5. $y = c_1 x^2 + c_2 x + c_3 x^{-1} - \frac{1}{12} x^{-2}$
7. $y = (c_1 + c_2 x + c_3 x^2)e^{3x} - \frac{1}{4}(\cos 3x - \sin 3x)$
9. $y = \cos x + \frac{1}{2} \sin 4x$
11. $y = e^{-3x}(-1.4 \cos x - \sin x)$
13. $y = 2 - 2 \sin x + \cos x$

Chapter 3 Review Questions and Problems, page 122

7. $y = c_1 + e^{-2x}(A \cos 3x + B \sin 3x)$
9. $y = c_1 \cosh 2x + c_2 \sinh 2x + c_3 \cos 2x + c_4 \sin 2x + \cosh x$
11. $y = (c_1 + c_2 x + c_3 x^2)e^{-1.5x}$
13. $y = (c_1 + c_2 x + c_3 x^2)e^{-2x} + x^2 - 3x + 3$
15. $y = c_1 x + c_2 x^{1/2} + c_3 x^{3/2} - \frac{10}{3}$
17. $y = 2e^{-2x} \cos 4x + 0.05\, x - 0.06$
19. $y = 4e^{-4x} + 5e^{-5x}$

Problem Set 4.1, page 136

1. Yes
5. $y_1' = 0.02(-y_1 + y_2)$, $y_2' = 0.02(y_1 - 2y_2 + y_3)$, $y_3' = 0.02(y_2 - y_3)$
7. $c_1 = 1$, $c_2 = -5$
9. $c_1 = 10$, $c_2 = 5$
11. $y_1' = y_2$, $y_2' = y_1 + \ + \frac{15}{4} y_2$, $\mathbf{y} = c_1[1 \quad 4]^\mathsf{T} e^{4t} + c_2 [1 \quad -\frac{1}{4}]^\mathsf{T} e^{-t/4}$
13. $y_1' = y_2$, $y_2' = 24y_1 - 2y_2$, $y_1 = c_1 e^{4t} + c_2 e^{-6t} = y$, $y_2 = y'$
15. **(a)** For example, $C = 1000$ gives -2.39993, -0.000167. **(b)** -2.4, 0.
(d) $a_{22} = -4 + 2\sqrt{6.4} = 1.05964$ gives the critical case. C about 0.18506.

Problem Set 4.3, page 147

1. $y_1 = c_1e^{-2t} + c_2e^{2t}, \quad y_2 = -3c_1e^{-2t} + c_2e^{2t}$
3. $y_1 = 2c_1e^{2t} + 2c_2, \quad y_2 = c_1e^{2t} - c_2$
5. $y_1 = 5c_1 + 2c_2e^{14.5t}$
$y_2 = -2c_1 + 5c_2e^{14.5t}$
7. $y_1 = -c_2 \cos \sqrt{2}t + c_3 \sin \sqrt{2}t + c_1$
$y_2 = c_2\sqrt{2} \sin \sqrt{2}t + c_3\sqrt{2} \cos \sqrt{2}t$
$y_3 = c_2 \cos \sqrt{2}t - c_3 \sin \sqrt{2}t + c_1$
9. $y_1 = \frac{1}{2}c_1e^{-18t} + 2c_2e^{9t} - c_3e^{18t}$
$y_2 = c_1e^{-18t} + c_2e^{9t} + c_3e^{18t}$
$y_3 = c_1e^{-18t} - 2c_2e^{9t} - \frac{1}{2}c_3e^{18t}$
11. $y_1 = -20e^t + 8e^{-t/2}$
$y_2 = 4e^t - 4e^{-t/2}$
13. $y_1 = 2 \sinh t, \quad y_2 = 2 \cosh t$
15. $y_1 = \frac{1}{2}e^t$
$y_2 = \frac{1}{2}e^t$
17. $y_2 = y_1' + y_1, \quad y_2' = y_1'' + y_1' = -y_1 - y_2 = -y_1 - (y_1' + y_1),$
$y_1'' + 2y_1' + 2y_1 = 0, \quad y_1 = e^{-t}(A \cos t + B \sin t),$
$y_2 = y_1' + y_1 = e^{-t}(B \cos t - A \sin t)$. Note that $r^2 = y_1^2 + y_2^2 = e^{-2t}(A^2 + B^2)$.
19. $I_1 = c_1e^{-t} + 3c_2e^{-3t}, I_2 = -3c_1e^{-t} - c_2e^{-3t}$

Problem Set 4.4, page 151

1. Unstable improper node, $y_1 = c_1e^t, \quad y_2 = c_2e^{2t}$
3. Center, always stable, $y_1 = A \cos 3t + B \sin 3t, \quad y_2 = 3B \cos 3t - 3A \sin 3t$
5. Stable spiral, $y_1 = e^{-2t}(A \cos 2t + B \sin 2t), \quad y_2 = e^{-2t}(B \cos 2t - A \sin 2t)$
7. Saddle point, always unstable, $y_1 = c_1e^{-t} + c_2e^{3t}, \quad y_2 = -c_1e^{-t} + c_2e^{3t}$
9. Unstable node, $y_1 = c_1e^{6t} + c_2e^{2t}, \quad y_2 = 2c_1e^{6t} - 2c_2e^{2t}$
11. $y = e^{-t}(A \cos t + B \sin t)$. Stable and attractive spirals
15. $p = 0.2 \neq 0$ (was 0), $\quad \Delta < 0$, spiral point, unstable.
17. For instance, **(a)** -2, **(b)** -1, **(c)** $= -\frac{1}{2}$, **(d)** $=1$, **(e)** 4.

Problem Set 4.5, page 159

5. Center at (0, 0). At (2, 0) set $y_1 = 2 + \tilde{y}_1$. Then $\tilde{y}_2' = \tilde{y}_1$. Saddle point at (2, 0).
7. (0, 0), $y_1' = -y_1 + y_2, \quad y_2' = -y_1 - y_2$, stable and attractive spiral point; $(-2, 2)$,
$y_1 = -2 + \tilde{y}_1, \quad y_2 = 2 + \tilde{y}_2, \quad \tilde{y}_1' = -\tilde{y}_1 - 3\tilde{y}_2, \quad \tilde{y}_2' = -\tilde{y}_1 - \tilde{y}_2$, saddle point
9. (0, 0) saddle point, $(-3, 0)$ and (3, 0) centers
11. $(\frac{1}{2}\pi \pm 2n\pi, 0)$ saddle points; $(-\frac{1}{2}\pi \pm 2n\pi, 0)$ centers.
Use $-\cos(\pm\frac{1}{2}\pi + \tilde{y}_1) = \sin(\pm\tilde{y}_1) \approx \pm\tilde{y}_1$.
13. $(\pm 2n\pi, 0)$ centers; $y_1 = (2n + 1)\pi + \tilde{y}_1', \quad (\pi \pm 2n\pi, 0)$ saddle points
15. By multiplication, $y_2y_2' = (4y_1 - y_1^3)y_1'$. By integration,
$y_2^2 = 4y_1^2 - \frac{1}{2}y_1^4 + c^* = \frac{1}{2}(c + 4 - y_1^2)(c - 4 + y_1^2)$, where $c^* = \frac{1}{2}c^2 - 8$.

Problem Set 4.6, page 163

3. $y_1 = c_1e^{-t} + c_2e^t, \quad y_2 = -c_1e^{-t} + c_2e^t - e^{3t}$
5. $y_1 = c_1e^{5t} + c_2e^{2t} - 0.43t - 0.24, \quad y_2 = c_1e^{5t} - 2c_2e^{2t} + 1.12t + 0.53$

7. $y_1 = c_1e^t + 4c_2e^{2t} - 3t - 4 - 2e^{-t}, \quad y_2 = -c_1e^t - 5c_2e^{2t} + 5t + 7.5 + e^{-t}$

9. The formula for $\mathbf{v}$ shows that these various choices differ by multiples of the eigenvector for $\lambda = -2$, which can be absorbed into, or taken out of, c_1 in the general solution $y^{(h)}$.

11. $y_1 = -\frac{8}{3}\cosh t - \frac{4}{3}\sinh t + \frac{11}{3}e^{2t}, \quad y_2 = -\frac{8}{3}\sinh t - \frac{4}{3}\cosh t + \frac{4}{3}e^{2t}$

13. $y_1 = \cos 2t + \sin 2t + 4\cos t, \; y_2 = 2\cos 2t - 2\sin 2t + \sin t$

15. $y_1 = 4e^{-t} - 4e^t + e^{2t}, \quad y_2 = -4e^{-t} + t$

17. $I_1 = 2c_1e^{\lambda_1 t} + 2c_2e^{\lambda_2 t} + 100,$
$I_2 = (1.1 + \sqrt{0.41})c_1e^{\lambda_1 t} + (1.1 - \sqrt{0.41})c_2e^{\lambda_2 t},$
$\lambda_1 = -0.9 + \sqrt{0.41}, \quad \lambda_2 = -0.9 - \sqrt{0.41}$

19. $c_1 = 17.948, \quad c_2 = -67.948$

Chapter 4 Review Questions and Problems, page 164

11. $y_1 = c_1e^{4t} + c_2e^{-4t}, \quad y_2 = 2c_1e^{4t} - 2c_2e^{-4t}$. Saddle point

13. $y_1 = e^{-4t}(A\cos t + B\sin t), \quad y_2 = \frac{1}{5}e^{-4t}[(B - 2A)\cos t - (A + 2B)\sin t]$; asymptotically stable spiral point

15. $y_1 = c_1e^{-5t} + c_2e^{-t}, \quad y_2 = c_1e^{-5t} - c_2e^{-t}$. Stable node

17. $y_1 = e^{-t}(A\cos 2t + B\sin 2t), \quad y_2 = e^{-t}(B\cos 2t - A\sin 2t)$. Stable and attractive spiral point

19. Unstable spiral point

21. $y_1 = c_1e^{-4t} + c_2e^{4t} - 1 - 8t^2, \quad y_2 = -c_1e^{-4t} + c_2e^{4t} - 4t$

23. $y_1 = 2c_1e^{-t} + 2c_2e^{3t} + \cos t - \sin t, \quad y_2 = -c_1e^{-t} + c_2e^{3t}$

25. $I_1' + 2.5(I_1 - I_2) = 169\sin t, \quad 2.5(I_2' - I_1') + 25I_2 = 0,$
$I_1 = (19 + 32.5t)e^{-5t} - 19\cos t + 62.5\sin t,$
$I_2 = (-6 - 32.5t)e^{-5t} + 6\cos t + 2.5\sin t$

27. (0, 0) saddle point; (−1, 0), (1, 0) centers

29. $(n\pi, 0)$ center when n is even and saddle point when n is odd

Problem Set 5.1, page 174

3. $\sqrt{|k|}$

5. $\sqrt{3/2}$

7. $y = a_0(1 - x^2 + x^4/2! - x^6/3! + - \cdots) = a_0e^{-x^2}$

9. $y = a_0 + a_1x - \frac{1}{2}a_0x^2 - \frac{1}{6}a_1x^3 + \cdots = a_0\cos x + a_1\sin x$

11. $a_0(1 - \frac{1}{12}x^4 - \frac{1}{60}x^5 - \cdots) + a_1(x + \frac{1}{2}x^2 + \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{24}x^5 - \cdots)$

13. $a_0(1 - \frac{1}{2}x^2 - \frac{1}{24}x^4 + \frac{13}{720}x^6 + \cdots) + a_1(x - \frac{1}{6}x^3 - \frac{1}{24}x^5 + \frac{5}{1008}x^7 + \cdots)$

15. $\displaystyle\sum_{m=1}^{\infty} \frac{(m+1)(m+2)}{(m+1)^2+1}x^m, \quad \sum_{m=5}^{\infty} \frac{(m-4)^2}{(m-3)!}x^m$

17. $s = 1 + x - x^2 - \frac{5}{6}x^3 + \frac{2}{3}x^4 + \frac{11}{24}x^5, \quad s(\frac{1}{2}) = \frac{923}{768}$

19. $s = 4 - x^2 - \frac{1}{3}x^3 + \frac{1}{30}x^5, \quad s(2) = -\frac{8}{5}$; but $x = 2$ is too large to give good values. Exact: $y = (x - 2)^2e^x$

Problem Set 5.2, page 179

5. $P_6(x) = \frac{1}{16}(231x^6 - 315x^4 + 105x^2 - 5),$
$P_7(x) = \frac{1}{16}(429x^7 - 693x^5 + 315x^3 - 35x)$

11. Set $x = az$. $y = c_1 P_n(x/a) + c_2 Q_n(x/a)$

15. $P_1^1 = \sqrt{1 - x^2}$, $P_2^1 = 3x\sqrt{1 - x^2}$, $P_2^2 = 3(1 - x^2)$,
$P_4^2 = (1 - x^2)(105x^2 - 15)/2$

Problem Set 5.3, page 186

3. $y_1 = 1 - \dfrac{x^2}{3!} + \dfrac{x^4}{5!} - + \cdots = \dfrac{\sin x}{x}$, $y_2 = \dfrac{1}{x} - \dfrac{x}{2!} + \dfrac{x^3}{4!} - + \cdots = \dfrac{\cos x}{x}$

5. $b_0 = 1$, $c_0 = 0$, $r^2 = 0$, $y_1 = e^{-x}$, $y_2 = e^{-x} \ln x$

7. $y_1 = 1 + \frac{1}{2}x^2 - \frac{1}{6}x^3 + \frac{1}{24}x^4 - \frac{1}{30}x^5 + \frac{1}{144}x^6 - \cdots$,
$y_2 = x + \frac{1}{6}x^3 - \frac{1}{12}x^4 + \frac{1}{120}x^5 - \frac{1}{120}x^6 + \cdots$

9. $y_1 \sqrt{x}$, $y_2 = 1 + x$

11. $y_1 = e^x$, $y_2 = e^x/x$

13. $y_1 = e^x$, $y_2 = e^x \ln x$

15. $y = AF(1, 1, -\frac{1}{2}; x) + Bx^{3/2}F(\frac{5}{2}, \frac{5}{2}, \frac{5}{2}; x)$

17. $y = A(1 - 8x + \frac{32}{5}x^2) + Bx^{3/4}F(\frac{7}{4}, -\frac{5}{4}, \frac{7}{4}; x)$

19. $y = c_1 F(2, -2, -\frac{1}{2}; t - 2) + c_2(t - 2)^{3/2}F(\frac{7}{2}, -\frac{1}{2}, \frac{5}{2}; t - 2)$

Problem Set 5.4, page 195

3. $c_1 J_0(\sqrt{x})$

5. $c_1 J_\nu(\lambda x) + c_2 J_{-\nu}(\lambda x)$, $\nu \neq 0, \pm 1, \pm 2, \cdots$

7. $c_1 J_{1/2}(\frac{1}{2}x) + c_2 J_{-1/2}(\frac{1}{2}x) = x^{-1/2}(\tilde{c}_1 \sin \frac{1}{2}x + \tilde{c}_2 \cos \frac{1}{2}x)$

9. $x^{-\nu}(c_1 J_\nu(x) + c_2 J_{-\nu}(x))$, $\nu \neq 0, \pm 1, \pm 2, \cdots$

13. $J_n(x_1) = J_n(x_2) = 0$ implies $x_1^{-n}J_n(x_1) = x_2^{-n}J_n(x_2) = 0$ and $[x^{-n}J_n(x)]' = 0$ somewhere between x_1 and x_2 by Rolle's theorem. Now use (21b) to get $J_{n+1}(x) = 0$ there. Conversely, $J_{n+1}(x_3) = J_{n+1}(x_4) = 0$, thus $x_3^{n+1}J_{n+1}(x_3) = x_4^{n+1}J_{n+1}(x_4) = 0$ implies $J_n(x) = 0$ in between by Rolle's theorem and (21a) with $\nu = n + 1$.

15. By Rolle, $J_0' = 0$ at least once between two zeros of J_0. Use $J_0' = -J_1$ by (21b) with $\nu = 0$. Together $J_1 = 0$ at least once between two zeros of J_0. Also use $(xJ_1)' = xJ_0$ by (21a) with $\nu = 1$ and Rolle.

19. Use (21b) with $\nu = 0$, (21a) with $\nu = 1$, (21d) with $\nu = 2$, respectively.

21. Integrate (21a).

23. Use (21a) with $\nu = 1$, partial integration, (21b) with $\nu = 0$, partial integration.

25. Use (21d) to get

$$\int J_5(x)\,dx = -2J_4(x) + \int J_3(x)\,dx = -2J_4(x) - 2J_2(x) + \int J_1(x)\,dx$$
$$= -2J_4(x) - 2J_2(x) - J_0(x) + c.$$

Problem Set 5.5, page 200

1. $c_1 J_4(x) + c_2 Y_4(x)$

3. $c_1 J_{2/3}(x^2) + c_2 Y_{2/3}(x^2)$

5. $c_1 J_0(\sqrt{x}) + c_2 Y_0(\sqrt{x})$

7. $\sqrt{x}\,(c_1 J_{1/4}(\frac{1}{2}kx^2) + c_2 Y_{1/4}(\frac{1}{2}kx^2))$
9. $x^3(c_1 J_3(x) + c_2 Y_3(x))$
11. Set $H^{(1)} = kH^{(2)}$ and use (10).
13. Use (20) in Sec. 5.4.

Chapter 5 Review Questions and Problems, page 200

11. $\cos 2x$, $\sin 2x$
13. $(x-1)^{-5}$, $(x-1)^7$; Euler–Cauchy with $x-1$ instead of x
15. $J_{\sqrt{5}}(x)$, $J_{-\sqrt{5}}(x)$
17. e^x, $1 + x$
19. $\sqrt{x}\,J_1(\sqrt{x})$, $\sqrt{x}\,Y_1(\sqrt{x})$

Problem Set 6.1, page 210

1. $3/s^2 + 12/s$
3. $s/(s^2 + \pi^2)$
5. $1/((s-2)^2 - 1)$
7. $(\omega \cos\theta + s \sin\theta)/(s^2 + \omega^2)$
9. $\dfrac{1}{s} + \dfrac{e^{-s} - 1}{s^2}$
11. $\dfrac{1 - e^{-bs}}{s^2} - \dfrac{be^{-bs}}{s}$
13. $\dfrac{(1 - e^{-s})^2}{s}$
15. $\dfrac{e^{-s} - 1}{2s^2} - \dfrac{e^{-s}}{2s} + \dfrac{1}{s}$
19. Use $e^{at} = \cosh at + \sinh at$.
23. Set $ct = p$. Then $\mathscr{L}(f(ct)) = \int_0^\infty e^{-st} f(ct)\,dt = \int_0^\infty e^{-(s/c)p} f(p)\,dp/c = F(s/c)/c$.
25. $0.2 \cos 1.8t + \sin 1.8t$
27. $\dfrac{1}{L^2} \cos \dfrac{n\pi t}{L}$
29. $2t^3 - 1.9t^5$
31. $\mathscr{L}^{-1}\left(\dfrac{4}{s-2} - \dfrac{3}{s+1}\right) = 4e^{2t} - 3e^{-t}$
33. $\dfrac{2}{(s+3)^3}$
35. $\dfrac{0.5 \cdot 2\pi}{(s+4.5)^2 + 4\pi^2}$
37. $\pi t e^{-\pi t}$
39. $\frac{7}{2} t^3 e^{-t\sqrt{2}}$
41. $e^{-5\pi t} \sinh \pi t$
43. $e^{3t}(2 \cos 3t + \frac{5}{3} \sin 3t)$
45. $(k_0 + k_1 t)e^{-at}$

Problem Set 6.2, page 216

1. $y = 1.25e^{-5.2t} - 1.25 \cos 2t + 3.25 \sin 2t$
3. $(s-3)(s+2) = 11s + 28 - 11 = 11s + 17$, $\quad Y = 10/(s-3) + 1/(s+2)$,
$y = 10e^{3t} + e^{-2t}$
5. $(s^2 - \frac{1}{4})Y = 12s$, $\quad y = 12 \cosh \frac{1}{2}t$
7. $y = \frac{1}{2}e^{3t} + \frac{5}{2}e^{-4t} + \frac{1}{2}e^{-3t}$
9. $y = e^t - e^{3t} + 2t$
11. $(s + 1.5)^2 Y = s + 31.5 + 3 + 54/s^4 + 64/s$,
$Y = 1/(s+1.5) + 1/(s+1.5)^2 + 24/s^4 - 32/s^3 + 32/s^2$,
$y = (1+t)e^{-1.5t} + 4t^3 - 16t^2 + 32t$
13. $t = \tilde{t} - 1$, $\quad \tilde{Y} = 4/(s-6)$, $\quad \tilde{y} = 4e^{6t}$, $\quad y = 4e^{6(t+1)}$

15. $t = \tilde{t} + 1.5,\quad (s-1)(s+4)\tilde{Y} = 4s + 17 + 6/(s-2),\quad y = 3e^{t-1.5} + e^{2(t-1.5)}$

17. $\dfrac{1}{(s+a)^2}$

19. $\dfrac{2\omega^2}{s(s^2+4\omega^2)}$

21. $\mathscr{L}(f') = \mathscr{L}(\sinh 2t) = s\mathscr{L}(f) - 1.$ *Answer:* $(s^2-2)/(s^3-4s)$

23. $12(1 - e^{-t/4})$

25. $(1 - \cos \omega t)/\omega^2$

27. $\frac{1}{9}(1 + t - \cos 3t - \frac{1}{3}\sin 3t)$

29. $\dfrac{1}{a^2}(e^{-at} - 1) + \dfrac{t}{a}$

Problem Set 6.3, page 223

3. $\mathscr{L}((t-2)u(t-2)) = e^{-2s}/s^2$

5. $\left(e^t\left(1 - u\left(t - \dfrac{1}{2}\pi\right)\right)\right) = \dfrac{1}{s-1}(1 - e^{-\pi s/2 + \pi/2})$

7. $\dfrac{1}{s+\pi}(e^{-2(s+\pi)} - e^{-4(s+\pi)})$

9. $e^{-3s/2}\left(\dfrac{2}{s^3} + \dfrac{3}{s^2} + \dfrac{\frac{9}{4}}{s}\right)$

11. $(se^{-\pi s/2} + e^{-\pi s})/(s^2+1)$

13. $2[1 + u(t-\pi)]\sin 3t$

15. $(t-3)^3 u(t-3)/6$

17. $e^{-t}\cos t\ (0 < t < 2\pi)$

19. $\frac{1}{3}(e^t - 1)^3 e^{-5t}$

21. $\sin 3t + \sin t\ (0 < t < \pi);\ \frac{4}{3}\sin 3t\ (t > \pi)$

23. $e^t - \sin t\ (0 < t < 2\pi),\quad e^t - \frac{1}{2}\sin 2t\ (t > 2\pi)$

25. $t - \sin t\ (0 < t < 1),\quad \cos(t-1) + \sin(t-1) - \sin t\ (t > 1)$

27. $t = 1 + \tilde{t},\quad \tilde{y}'' + 4\tilde{y} = 8(1+\tilde{t})^2(1 - u(\tilde{t} - 4)),\quad \cos 2t + 2t^2 - 1$ if $t < 5$,
$\cos 2t + 49\cos(2t - 10) + 10\sin(2t - 10)$ if $t > 5$

29. $0.1i' + 25i = 490e^{-5t}[1 - u(t-1)]$,
$i = 20(e^{-5t} - e^{-250t}) + 20u(t-1)[-e^{-5t} + e^{-250t+245}]$

31. $Rq' + q/C = 0,\quad Q = \mathscr{L}(q),\quad q(0) = CV_0,\quad i = q'(t)$,
$R(sQ - CV_0) + Q/C = 0,\quad q = CV_0 e^{-t/(RC)}$

33. $10I + \dfrac{100}{s}I = \dfrac{100}{s^2}e^{-2s},\quad I = e^{-2s}\left(\dfrac{1}{s} - \dfrac{1}{s+10}\right),\quad i = 0$ if $t < 2$ and
$1 - e^{-10(t-2)}$ if $t > 2$

35. $i = (10\sin 10t + 100\sin t)(u(t-\pi) - u(t-3\pi))$

37. $(0.5s^2 + 20)I = 78s(1 + e^{-\pi s})/(s^2+1)$,
$i = 4\cos t - 4\cos\sqrt{40}t - 4u(t-\pi)[\cos t + \cos(\sqrt{40}(t-\pi))]$

39. $i' + 2i + 2\displaystyle\int_0^t i(\tau)\,d\tau = 1000(1 - u(t-2)),\quad I = 1000(1 - e^{-2s})/(s^2 + 2s + 2)$,

$i = 1000e^{-t}\sin t - 1000u(t-2)e^{-t+2}\sin(t-2)$

Problem Set 6.4, page 230

3. $y = 8\cos 2t + \frac{1}{2}u(t-\pi)\sin 2t$

5. $\sin t\ (0 < t < \pi);\quad 0\ (\pi < t < 2\pi);\quad -\sin t\ (t > 2\pi)$

7. $y = e^{-t} + 4e^{-3t}\sin\frac{1}{2}t + \frac{1}{2}u(t - \frac{1}{2})e^{-3(t-1/2)}\sin(\frac{1}{2}t - \frac{1}{4})$

9. $y = 0.1[e^t + e^{-2t}(-\cos t + 7\sin t)] + 0.1u(t-10)[-e^{-t} +$
$e^{-2t+30}(\cos(t-10) - 7\sin(t-10))]$

11. $y = -e^{-3t} + e^{-2t} + \frac{1}{6}u(t-1)(1 - 3e^{-2(t-1)} + 2e^{-3(t-1)}) + u(t-2)(e^{-2(t-2)} - e^{-3(t-2)})$

15. $ke^{-ps}/(s - se^{-ps})\ (s > 0)$

Problem Set 6.5, page 237

1. t

3. $(e^t - e^{-t})/2 = \sinh t$

5. $\frac{1}{2}t \sin \omega t$

7. $e^t - t - 1$

9. $y - 1 * y = 1, \quad y = e^t$

11. $y = \cos t$

13. $y(t) + 2\int_0^t e^{t-\tau} y(\tau)\, d\tau = te^t, \quad y = \sinh t$

17. $e^{4t} - e^{-1.5t}$

19. $t \sin \pi t$

21. $(\omega t - \sin \omega t)/\omega^2$

23. $4.5(\cosh 3t - 1)$

25. $1.5t \sin 6t$

Problem Set 6.6, page 241

3. $\dfrac{\frac{1}{2}}{(s+3)^2}$

5. $\dfrac{s^2 - \omega^2}{(s^2 + \omega^2)^2}$

7. $\dfrac{2s^3 + 24s}{(s^2 - 4)^3}$

9. $\dfrac{\pi(3s^2 - \pi^2)}{(s^2 + \pi^2)^3}$

11. $\dfrac{4s^2 - \pi^2}{(s^2 + \frac{1}{4}\pi^2)^2}$

15. $F(s) = -\dfrac{1}{2}\left(\dfrac{1}{s^2 - 9}\right)', \quad f(t) = \frac{1}{6}t \sinh 3t$

17. $\ln s - \ln(s-1); \quad (-1 + e^t)/t$

19. $[\ln(s^2 + 1) - 2\ln(s-1)]' = 2s/(s^2+1) - 2/(s-1); \quad 2(-\cos t + e^t)/t$

Problem Set 6.7, page 246

3. $y_1 = -e^{-5t} + 4e^{2t}, \quad y_2 = e^{-5t} + 3e^{2t}$

5. $y_1 = -\cos t + \sin t + 1 + u(t-1)[-1 + \cos(t-1) - \sin(t-1)]$
$y_2 = \cos t + \sin t - 1 + u(t-1)[1 - \cos(t-1) - \sin(t-1)]$

7. $y_1 = -e^{-2t} + 4e^t + \frac{1}{3}u(t-1)(-e^{3-2t} + e^t),$
$y_2 = -e^{-2t} + e^t + \frac{1}{3}u(t-1)(-e^{3-2t} + e^t)$

9. $y_1 = (3 + 4t)e^{3t}, \quad y_2 = (1 - 4t)e^{3t}$

11. $y_1 = e^t + e^{2t}, \quad y_2 = e^{2t}$

13. $y_1 = -4e^t + \sin 10t + 4\cos t, \quad y_2 = 4e^t - \sin 10t + 4\cos t$

15. $y_1 = e^t, \quad y_2 = e^{-t}, \quad y_3 = e^t - e^{-t}$

19. $4i_1 + 8(i_1 - i_2) + 2i_1' = 390\cos t, \quad 8i_2 + 8(i_2 - i_1) + 4i_2' = 0,$
$i_1 = -26e^{-2t} - 16e^{-8t} + 42\cos t + 15\sin t,$
$i_2 = -26e^{-2t} + 8e^{-8t} + 18\cos t + 12\sin t$

Chapter 6 Review Questions and Problems, page 251

11. $\dfrac{5s}{s^2 - 4} - \dfrac{3}{s^2 - 1}$

13. $\frac{1}{2}(1 - \cos \pi t), \quad \pi^2/(2s^3 + 2\pi^2 s)$

15. $e^{-3s+3/2}/(s - \frac{1}{2})$

17. Sec. 6.6; $2s^2/(s^2+1)^2$

19. $12/(s^2(s+3))$

21. $tu(t-1)$

23. $\sin(\omega t + \theta)$

25. $3t^2 + t^3$

27. $e^{-2t}(3\cos t - 2\sin t)$

29. $y = e^{-2t}(13\cos t + 11\sin t) + 10t - 8$

31. $e^{-t} + u(t-\pi)[1.2\cos t - 3.6\sin t + 2e^{-t+\pi} - 0.8e^{2t-2\pi}]$

33. $0\ (0 \leqq t \leqq 2), \quad 1 - 2e^{-(t-2)} + e^{-2(t-2)} \quad (t > 2)$

35. $y_1 = 4e^t - e^{-2t}, \quad y_2 = e^t - e^{-2t}$

37. $y_1 = \cos t - u(t-\pi)\sin t + 2u(t-2\pi)\sin^2 \frac{1}{2}t,$
$y_2 = -\sin t - 2u(t-\pi)\cos^2 \frac{1}{2}t + u(t-2\pi)\sin t$

39. $y_1 = (1/\sqrt{10})\sin\sqrt{10}t, \quad y_2 = -(1/\sqrt{10})\sin\sqrt{10}t$

41. $1 - e^{-t}\ (0 < t < 4), \quad (e^4 - 1)e^{-t} \quad (t > 4)$

43. $i(t) = e^{-4t}(\frac{3}{26}\cos 3t - \frac{10}{39}\sin 3t) - \frac{3}{26}\cos 10t + \frac{8}{65}\sin 10t$

45. $5i_1' + 20(i_1 - i_2) = 60, \quad 30i_2' + 20(i_2' - i_1') + 20i_2 = 0,$
$i_1 = -8e^{-2t} + 5e^{-0.8t} + 3, \quad i_2 = -4e^{-2t} + 4e^{-0.8t}$

Problem Set 7.1, page 261

3. $3 \times 3, \quad 3 \times 4, \quad 3 \times 6, \quad 2 \times 2, \quad 2 \times 3, \quad 3 \times 2$

5. $\mathbf{B} = \frac{1}{5}\mathbf{A}, \quad \frac{1}{10}\mathbf{A}$

7. No, no, yes, no, no

9. $\begin{bmatrix} 0 & 6 & 12 \\ 18 & 15 & 15 \\ 3 & 0 & -9 \end{bmatrix}, \begin{bmatrix} 0 & 2.5 & 1 \\ 2.5 & 1.5 & 2 \\ -1 & 2 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 8.5 & 13 \\ 20.5 & 16.5 & 17 \\ 2 & 2 & -10 \end{bmatrix}$, undefined

11. $\begin{bmatrix} 0 & 26 \\ 34 & 32 \\ 28 & -10 \end{bmatrix}$, same, $\begin{bmatrix} 5.4 & 0.6 \\ -4.2 & 2.4 \\ -0.6 & 0.6 \end{bmatrix}$, same

13. $\begin{bmatrix} 70 & 28 \\ -28 & 56 \\ 14 & 0 \end{bmatrix}$, same, $-\mathbf{D}$, undefined

15. $\begin{bmatrix} 5.5 \\ 33.0 \\ -11.0 \end{bmatrix}$, same, undefined, undefined

17. $\begin{bmatrix} -4.5 \\ -27.0 \\ 9.0 \end{bmatrix}$

Problem Set 7.2, page 270

5. $10, \ n(n+1)/2$

7. $\mathbf{0}, \quad \mathbf{I}, \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} 10 & -14 & -6 \\ -5 & 7 & -12 \\ -5 & -1 & -4 \end{bmatrix}$, same, $\begin{bmatrix} 10 & -5 & -15 \\ -14 & 7 & -33 \\ -2 & -4 & -4 \end{bmatrix}$, same

13. $\begin{bmatrix} 1 & 2 & 0 \\ 2 & 13 & -6 \\ 0 & -6 & 4 \end{bmatrix}$, $\begin{bmatrix} -9 & -5 \\ 3 & -1 \\ 4 & 0 \end{bmatrix}$, undefined, $\begin{bmatrix} -9 & 3 & 4 \\ -5 & -1 & 0 \end{bmatrix}$

15. Undefined, $\begin{bmatrix} 8 \\ -4 \\ -3 \end{bmatrix}$, $[7 \quad -1 \quad 3]$, same

17. $\begin{bmatrix} -30 & -18 \\ 45 & 9 \\ 5 & -7 \end{bmatrix}$, undefined, $\begin{bmatrix} 22 \\ 4 \\ -12 \end{bmatrix}$, undefined

19. Undefined, $\begin{bmatrix} 10.5 \\ 0 \\ -3 \end{bmatrix}$, $\begin{bmatrix} 7 \\ -3 \\ 1 \end{bmatrix}$, same

25. **(d)** $\mathbf{AB} = (\mathbf{AB})^\mathsf{T} = \mathbf{B}^\mathsf{T}\mathbf{A}^\mathsf{T} = \mathbf{BA}$; etc.
(e) *Answer.* If $\mathbf{AB} = -\mathbf{BA}$.

29. $\mathbf{p} = [85 \quad 62 \quad 30]^\mathsf{T}$, $\mathbf{v} = [44{,}920 \quad 30{,}940]^\mathsf{T}$

Problem Set 7.3, page 280

1. $x = -2, \quad y = 0.5$

3. $x = 1, \quad y = 3, \quad z = -5$

5. $x = 6, \quad y = -7$

7. $x = -3t, \quad y = t$ arb., $\quad z = 2t$

9. $x = 3t - 1, \quad y = -t + 4, \quad z = t$ arb.

11. $w = 1, \quad x = t_1$ arb., $\quad y = 2t_2 - t_1, \quad z = t_2$ arb.

13. $w = 4, \quad x = 0, \quad y = 2, \quad z = 6$

17. $I_1 = 2, \quad I_2 = 6, \quad I_3 = 8$

19. $I_1 = (R_1 + R_2)E_0/(R_1R_2)$ A, $\quad I_2 = E_0/R_1$ A, $\quad I_3 = E_0/R_2$ A

21. $x_2 = 1600 - x_1, \quad x_3 = 600 + x_1, \quad x_4 = 1000 - x_1$. No

23. C: $3x_1 - x_3 = 0$, H: $8x_1 - 2x_4 = 0$, O: $2x_2 - 2x_3 - x_4 = 0$, thus $C_3H_8 + 5O_2 \rightarrow 3CO_2 + 4H_2O$

Problem Set 7.4, page 287

1. 1; $[2 \quad -1 \quad 3]$; $[2 \quad -1]^\mathsf{T}$

3. 3; $\{[3 \quad 5 \quad 0], \ [0 \quad 3 \quad 5], \ [0 \quad 0 \quad 1]\}$

5. 3; $\{[2 \quad -1 \quad 4], \ [0 \quad 1 \quad -46], \ [0 \quad 0 \quad 1]\}$; $\{[2 \quad 0 \quad 1], \ [0 \quad 3 \quad 23], \ [0 \quad 0 \quad 1]\}$

7. 2; [8 0 4 0], [0 2 0 4]; [8 0 4], [0 2 0]
9. 3; [9 0 1 0], [0 9 8 9], [0 0 1 0]
11. **(c)** 1
17. No
19. Yes
21. No
23. Yes
25. Yes
27. 2, [−2 0 1], [0 2 1]
29. No
31. No
33. 1, solution of the given system $c[1 \quad \frac{10}{3} \quad 3]$, basis $[1 \quad \frac{10}{3} \quad 3]$
35. 1, $[4 \quad 2 \quad \frac{4}{3} \quad 1]$

Problem Set 7.7, page 300

7. $\cos(\alpha + \beta)$
9. 1
11. 40
13. 289
15. −64
17. 2
19. 2
21. $x = 3.5, \quad y = -1.0$
23. $x = 0, \quad y = 4, \quad z = -1$
25. $w = 3, \quad x = 0, \quad y = 2, \quad z = -2$

Problem Set 7.8, page 308

1. $\begin{bmatrix} 1.20 & 4.64 \\ 0.50 & 3.60 \end{bmatrix}$

3. $\begin{bmatrix} 54 & 0.9 & -3.4 \\ 2 & 0.2 & -0.2 \\ -30 & -0.5 & 2 \end{bmatrix}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 3 & -4 & 1 \end{bmatrix}$

7. $\mathbf{A}^{-1} = \mathbf{A}$

9. $\begin{bmatrix} 0 & 0 & \frac{1}{2} \\ \frac{1}{8} & 0 & 0 \\ 0 & \frac{1}{4} & 0 \end{bmatrix}$

11. $(\mathbf{A}^2)^{-1} = (\mathbf{A}^{-1})^2 = \begin{bmatrix} 3.760 & 22.272 \\ 2.400 & 15.280 \end{bmatrix}$

15. $\mathbf{AA}^{-1} = \mathbf{I}, \quad (\mathbf{AA}^{-1})^{-1} = (\mathbf{A}^{-1})^{-1}\mathbf{A}^{-1} = \mathbf{I}$. Multiply by $\mathbf{A}$ from the right.

Problem Set 7.9, page 318

1. $[1 \quad 0]^T, \quad [0 \quad 1]^T; \quad [1 \quad 0]^T, \quad [0 \quad -1]^T; \quad [1 \quad 1]^T, \quad [-1 \quad 1]^T$
3. 1, $[1 \quad 11 \quad -7]^T$
5. No
7. Dimension 2, basis xe^{-x}, e^{-x}
9. 3; basis $\begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}$
11. $x_1 = 5y_1 - y_2, \quad x_2 = 3y_1 - y_2$
13. $x_1 = 2y_1 - 3y_2, \quad x_2 = -10y_1 + 16y_2 + y_3, \quad x_3 = -7y_1 + 11y_2 + y_3$

15. $\sqrt{26}$

17. $\sqrt{5}$

19. 1

21. $k = -20$

23. $\mathbf{a} = [3 \quad 1 \quad -4]^T$, $\mathbf{b} = [-4 \quad 8 \quad -1]^T$, $\|\mathbf{a} + \mathbf{b}\| = \sqrt{107} \leqq 5.099 + 9$

25. $\mathbf{a} = [5 \quad 3 \quad 2]^T$, $\mathbf{b} = [3 \quad 2 \quad -1]^T$, $90 + 14 = 2(38 + 14)$

Chapter 7 Review Questions and Problems, page 318

11. $\begin{bmatrix} -1 & 6 & 1 \\ -18 & 8 & -7 \\ -13 & -2 & -7 \end{bmatrix}$, $\begin{bmatrix} 1 & 18 & 13 \\ -6 & -8 & 2 \\ -1 & 7 & 7 \end{bmatrix}$

13. $[21 \quad -8 \quad -31]^T$, $[21 \quad -8 \quad 31]$

15. 197, 0

17. -5, $\det \mathbf{A}^2 = (\det \mathbf{A})^2 = 25$, 0

19. $\begin{bmatrix} -2 & -12 & -12 \\ -12 & 16 & -9 \\ -12 & -9 & -14 \end{bmatrix}$

21. $x = 4$, $y = -2$, $z = 8$

23. $x = 6$, $y = 2t + 2$, $z = t$ arb.

25. $x = 0.4$, $y = -1.3$, $z = 1.7$

27. $x = 10$, $y = -2$

29. Ranks 2, 2, ∞

31. Ranks 2, 2, 1

33. $I_1 = 16.5$ A, $I_2 = 11$ A, $I_3 = 5.5$ A

35. $I_1 = 4$ A, $I_2 = 5$ A, $I_3 = 1$ A

Problem Set 8.1, page 329

1. 3, $[1 \quad 0]^T$; -0.6, $[0 \quad 1]^T$

3. -4, $[2 \quad 9]^T$; 3, $[1 \quad 1]^T$

5. $-3i$, $[1 \quad -i]$; $3i$, $[1 \quad i]$, $i = \sqrt{-1}$

7. $\lambda^2 = 0$, $[1 \quad 0]^T$

9. $0.8 + 0.6i$, $[1 \quad -i]^T$; $0.8 - 0.6i$, $[1 \quad i]^T$

11. $-(\lambda^3 - 18\lambda^2 + 99\lambda - 162)/(\lambda - 3) = -(\lambda^2 - 15\lambda + 54)$; 3, $[2 \quad -2 \quad 1]^T$; 6, $[1 \quad 2 \quad 2]^T$; 9, $[2 \quad 1 \quad -2]^T$

13. $-(\lambda - 9)^3$; 9, $[2 \quad -2 \quad 1]^T$, defect 2

15. $(\lambda + 1)^2(\lambda^2 + 2\lambda - 15)$; -1, $[1 \quad 0 \quad 0 \quad 0]^T$, $[0 \quad 1 \quad 0 \quad 0]^T$; -5, $[-3 \quad -3 \quad 1 \quad 1]^T$, 3, $[3 \quad -3 \quad 1 \quad -1]^T$

17. $\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$. Eigenvalues i, $-i$. Corresponding eigenvectors are complex, indicating that no direction is preserved under a rotation.

19. $\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$; 1, $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$; 0, $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$. A point onto the x_2-axis goes onto itself, a point on the x_1-axis onto the origin.

23. Use that real entries imply real coefficients of the characteristic polynomial.

Problem Set 8.2, page 333

1. 1.5, $[1 \quad -1]^\mathsf{T}$, $-45°$; 4.5, $[1 \quad 1]^\mathsf{T}$, $45°$
3. 1, $[-1/\sqrt{6} \quad 1]^\mathsf{T}$, 112.2°; 8, $[1 \quad 1/\sqrt{6}]^\mathsf{T}$, 22.2°
5. 0.5, $[1 \quad -1]^\mathsf{T}$; 1.5, $[1 \quad 1]^\mathsf{T}$; directions $-45°$ and 45°
7. $[5 \quad 8]^\mathsf{T}$
9. $[11 \quad 12 \quad 16]^\mathsf{T}$
11. 1.8
13. $c[10 \quad 18 \quad 25]^\mathsf{T}$
15. $\mathbf{x} = (\mathbf{I} - \mathbf{A})^{-1}\mathbf{y} = [0.6747 \quad 0.7128 \quad 0.7543]^\mathsf{T}$
17. $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j\ (\mathbf{x}_j \neq 0)$, $(\mathbf{A} - k\mathbf{I})\mathbf{x}_j = \lambda_j\mathbf{x}_j - k\mathbf{x}_j = (\lambda_j - k)\mathbf{x}_j$.
19. From $\mathbf{A}\mathbf{x}_j = \lambda_j\mathbf{x}_j\ (\mathbf{x}_j \neq \mathbf{0})$ and Prob. 18 follows $k_p\mathbf{A}^p\mathbf{x}_j = k_p\lambda_j^p\mathbf{x}_j$ and $k_q\mathbf{A}^q\mathbf{x}_j = k_q\lambda_j^q\mathbf{x}_j\ (p \geqq 0, q \geqq 0$, integer). Adding on both sides, we see that $k_p\mathbf{A}^p + k_q\mathbf{A}^q$ has the eigenvalue $k_p\lambda_j^p + k_q\lambda_j^q$. From this the statement follows.

Problem Set 8.3, page 338

1. $0.8 \pm 0.6i$, $[1 \quad \pm i]^\mathsf{T}$; orthogonal
3. $2 \pm 0.8i$, $[1 \quad \pm i]$. Not skew–symmetric!
5. 1, $[0 \quad 2 \quad 1]^\mathsf{T}$; 6, $[1 \quad 0 \quad 0]^\mathsf{T}$, $[0 \quad 1 \quad -2]^\mathsf{T}$; symmetric
7. 0, $\pm 25i$, skew–symmetric
9. 1, $[0 \quad 1 \quad 0]^\mathsf{T}$; i, $[1 \quad 0 \quad i]^\mathsf{T}$; $-i$, $[1 \quad 0 \quad -i]^\mathsf{T}$, orthogonal
15. No
17. $\mathbf{A}^{-1} = (-\mathbf{A}^\mathsf{T})^{-1} = -(\mathbf{A}^{-1})^\mathsf{T}$
19. No since $\det \mathbf{A} = \det(\mathbf{A}^\mathsf{T}) = \det(-\mathbf{A}) = (-1)^3 \det(\mathbf{A}) = -\det(\mathbf{A}) = 0$.

Problem Set 8.4, page 345

1. $\begin{bmatrix} -25 & 12 \\ -50 & 25 \end{bmatrix}$, -5, $\begin{bmatrix} 3 \\ 5 \end{bmatrix}$; 5, $\begin{bmatrix} 2 \\ 5 \end{bmatrix}$; $\mathbf{x} = \begin{bmatrix} -2 \\ 4 \end{bmatrix}$, $\begin{bmatrix} 2 \\ 1 \end{bmatrix}$

3. $\begin{bmatrix} 3.008 & -0.544 \\ 5.456 & 6.992 \end{bmatrix}$, 4, $\begin{bmatrix} -17 \\ 31 \end{bmatrix}$; 6, $\begin{bmatrix} -2 \\ 11 \end{bmatrix}$; $\mathbf{x} = \begin{bmatrix} 25 \\ 25 \end{bmatrix}$, $\begin{bmatrix} 10 \\ 5 \end{bmatrix}$

5. $\begin{bmatrix} 4 & 3 & -9 \\ 0 & -5 & 15 \\ 0 & -5 & 15 \end{bmatrix}$, 0, $\begin{bmatrix} 0 \\ 3 \\ 1 \end{bmatrix}$; 4, $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$; 10, $\begin{bmatrix} -1 \\ 1 \\ 1 \end{bmatrix}$; $\mathbf{x} = \begin{bmatrix} 3 \\ 0 \\ 1 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 1 \\ -1 \\ 1 \end{bmatrix}$

9. $\begin{bmatrix} \frac{1}{5} & \frac{2}{5} \\ -\frac{2}{5} & \frac{1}{5} \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} = \begin{bmatrix} 5 & 0 \\ 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} -2 & 1 \\ 3 & -1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 1 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & -5 \end{bmatrix}$

13. $\begin{bmatrix} 1 & 0 & 0 \\ -2 & 1 & 0 \\ 1 & -2 & 1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 3 & 2 & 1 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -2 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

15. $\begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ -\frac{1}{3} & \frac{1}{6} & \frac{1}{6} \\ 0 & -\frac{1}{2} & \frac{1}{2} \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & -2 & 0 \\ 1 & 1 & -1 \\ 1 & 1 & 1 \end{bmatrix} = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 5 \end{bmatrix}$

17. $\mathbf{C} = \begin{bmatrix} 7 & 3 \\ 3 & 7 \end{bmatrix}$, $4y_1^2 + 10y_2^2 = 200$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \mathbf{y}$, ellipse

19. $\mathbf{C} = \begin{bmatrix} 3 & 11 \\ 11 & 3 \end{bmatrix}$, $14y_1^2 - 8y_2^2 = 0$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{y}$; pair of straight lines

21. $\mathbf{C} = \begin{bmatrix} 1 & -6 \\ -6 & 1 \end{bmatrix}$, $7y_1^2 - 5y_2^2 = 70$, $\mathbf{x} = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 & 1 \\ 1 & 1 \end{bmatrix} \mathbf{y}$, hyperbola

23. $\mathbf{C} = \begin{bmatrix} -11 & 42 \\ 42 & 24 \end{bmatrix}$, $52y_1^2 - 39y_2^2 = 156$, $\mathbf{x} = \frac{1}{\sqrt{13}} \begin{bmatrix} 2 & 3 \\ 3 & -2 \end{bmatrix} \mathbf{y}$, hyperbola

Problem Set 8.5, page 351

1. Hermitian, 5, $[-i \quad 1]^{\mathsf{T}}$, 7, $[i \quad 1]^{\mathsf{T}}$
3. Unitary, $(1 - i\sqrt{3})/2$, $[-1 \quad 1]^{\mathsf{T}}$; $(1 + i\sqrt{3})/2$, $[1 \quad 1]^{\mathsf{T}}$
5. Skew-Hermitian, unitary, $-i$, $[0 \quad -1 \quad 1]^{\mathsf{T}}$, i, $[1 \quad 0 \quad 0]^{\mathsf{T}}$, $[0 \quad 1 \quad 1]^{\mathsf{T}}$
7. Eigenvalues -1, 1; eigenvectors $[1 \quad -1]^{\mathsf{T}}$, $[1 \quad 1]^{\mathsf{T}}$; $[1 \quad -i]^{\mathsf{T}}$, $[1 \quad i]^{\mathsf{T}}$; $[0 \quad 1]^{\mathsf{T}}$, $[1 \quad 0]^{\mathsf{T}}$, resp.
9. Hermitian, 16
11. Skew-Hermitian, $-6i$
13. $\overline{(\mathbf{ABC})}^{\mathsf{T}} = \overline{\mathbf{C}}^{\mathsf{T}}\overline{\mathbf{B}}^{\mathsf{T}}\overline{\mathbf{A}}^{\mathsf{T}} = \mathbf{C}^{-1}(-\mathbf{B})\mathbf{A}$
15. $\mathbf{A} = \mathbf{H} + \mathbf{S}$, $\mathbf{H} = \frac{1}{2}(\mathbf{A} + \overline{\mathbf{A}}^{\mathsf{T}})$, $\mathbf{S} = \frac{1}{2}(\mathbf{A} - \overline{\mathbf{A}}^{\mathsf{T}})$ (**H** Hermitian, **S** skew-Hermitian)
19. $\mathbf{A}\overline{\mathbf{A}}^{\mathsf{T}} - \overline{\mathbf{A}}^{\mathsf{T}}\mathbf{A} = (\mathbf{H} + \mathbf{S})(\mathbf{H} - \mathbf{S}) - (\mathbf{H} - \mathbf{S})(\mathbf{H} + \mathbf{S}) = 2(-\mathbf{HS} + \mathbf{SH}) = \mathbf{0}$ if and only if $\mathbf{HS} = \mathbf{SH}$.

Chapter 8 Review Questions and Problems, page 352

11. 3, $[1 \quad 1]^{\mathsf{T}}$; 2, $[1 \quad -1]^{\mathsf{T}}$
13. 3, $[1 \quad 5]^{\mathsf{T}}$; 7, $[1 \quad 1]^{\mathsf{T}}$
15. 0, $[2 \quad -2 \quad 1]^{\mathsf{T}}$; $9i$, $[-1 + 3i \quad 1 + 3i \quad 4]^{\mathsf{T}}$; $-9i$, $[-1 - 3i \quad 1 - 3i \quad 4]^{\mathsf{T}}$

17. $-1, 1$; $\mathbf{A} = \frac{1}{16} \begin{bmatrix} 5 & -3 \\ -3 & 5 \end{bmatrix} \begin{bmatrix} 23 & 2 \\ 39 & 1 \end{bmatrix} = \frac{1}{8} \begin{bmatrix} -1 & 1 \\ 63 & 1 \end{bmatrix}$

19. $\frac{1}{3}\begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \mathbf{A} \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} -0.9 & 0 \\ 0 & 0.6 \end{bmatrix}$

21. $\frac{1}{3}\begin{bmatrix} 1 & 1 & -1 \\ 1 & -1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \mathbf{A} \begin{bmatrix} 1 & 2 & 1 \\ 1 & -1 & 1 \\ -1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 4 & 0 & 0 \\ 0 & -20 & 0 \\ 0 & 0 & 22 \end{bmatrix}$

23. $\mathbf{C} = \begin{bmatrix} 4 & 12 \\ 12 & -14 \end{bmatrix}$, $10y_1^2 - 20y_2^2 = 20$, $\mathbf{x} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}\mathbf{y}$, hyperbola

25. $\mathbf{C} = \begin{bmatrix} 3.7 & 1.6 \\ 1.6 & 1.3 \end{bmatrix}$, $4.5y_1^2 + 0.5y_2^2 = 4.5$, $\mathbf{x} = \frac{1}{\sqrt{5}}\begin{bmatrix} 2 & 1 \\ 1 & -2 \end{bmatrix}\mathbf{y}$, ellipse

Problem Set 9.1, page 360

1. 5, 1, 0; $\sqrt{26}$; $[5/\sqrt{26}, 1/\sqrt{26}, 0]$

3. 8.5, -4.0, 1.7; $\sqrt{91.14}$, $[0.890, -0.419, 0.178]$

5. 2, 1, -2; $\mathbf{u} = [\frac{2}{3}, \frac{1}{3}, -\frac{2}{3}]$, position vector of Q

7. Q: $(4, 0, \frac{1}{2})$, $|\mathbf{v}| = \sqrt{16.25}$

9. Q: $(0, 0, -8)$, $|\mathbf{v}| = 8$

11. $[6, 4, 0]$, $[\frac{3}{2}, 1, 0]$, $[-3, -2, 0]$

13. $[1, 5, 8]$

15. $7[9, -7, 8] = [63, -49, 56]$

17. $[12, 8, 0]$

21. $[4, 9, -3]$, $\sqrt{106}$

23. $[0, 0, 5]$, 5

25. $[6, 2, -14] = 2\mathbf{u}$, $\sqrt{236}$

27. $\mathbf{p} = [0, 0, -5]$

29. $\mathbf{v} = [v_1, v_2, 3]$, v_1, v_2 arbitrary

31. $k = 10$

33. $|\mathbf{p} + \mathbf{q} + \mathbf{u}| \leqq 18$. Nothing

35. $v_B - v_A = [-19, 0] - [22/\sqrt{2}, 22/\sqrt{2}] = [-19 - 22/\sqrt{2}, -22/\sqrt{2}]$

37. $\mathbf{u} + \mathbf{v} + \mathbf{p} = [-k, 0] + [l, l] + [0, -1000] = \mathbf{0}$, $-k + l + 0 = 0$,
$0 + l - 1000 = 0$, $l = 1000, k = 1000$

Problem Set 9.2, page 367

1. 44, 44, 0

3. $\sqrt{35}$, $\sqrt{320}$, $\sqrt{86}$

5. $|[2, 9, 9]| = \sqrt{166} = 12.88 < \sqrt{80} + \sqrt{86} = 18.22$

7. $|-24| = 24$, $|a||c| = \sqrt{35}\sqrt{86} = \sqrt{3010} = 54.86$; cf. (6)

9. 300; cf. (5a) and (5b)

13. Use (1) and $|\cos\gamma| \leqq 1$.

15. $|\mathbf{a} + \mathbf{b}|^2 + |\mathbf{a} - \mathbf{b}|^2 = \mathbf{a} \cdot \mathbf{a} + 2\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b} + (\mathbf{a} \cdot \mathbf{a} - 2\mathbf{a} \cdot \mathbf{b} + \mathbf{b} \cdot \mathbf{b})$
$= 2|\mathbf{a}|^2 + 2|\mathbf{b}|^2$

17. $[2, 5, 0] \cdot [2, 2, 2] = 14$

19. $[0, 4, 3] \cdot [-3, -2, 1] = -5$ is negative! Why?

21. Yes, because $W = (\mathbf{p} + \mathbf{q}) \cdot \mathbf{d} = \mathbf{p} \cdot \mathbf{d} + \mathbf{q} \cdot \mathbf{d}$.

23. arccos $0.5976 = 53.3°$

27. $\beta - \alpha$ is the angle between the unit vectors **a** and **b**. Use (2).

29. $\gamma = \arccos(12/(6\sqrt{13})) = 0.9828 = 56.3°$ and $123.7°$

31. $a_1 = -\frac{28}{3}$

33. $\pm[\frac{3}{5}, -\frac{4}{5}]$

35. $(\mathbf{a} + \mathbf{b}) \cdot (\mathbf{a} - \mathbf{b}) = |\mathbf{a}|^2 - |\mathbf{b}|^2 = 0$, $|\mathbf{a}| = |\mathbf{b}|$. A square.

37. 0. Why?

39. If $|\mathbf{a}| = |\mathbf{b}|$ or if **a** and **b** are orthogonal.

Problem Set 9.3, page 374

5. $-\mathbf{m}$ instead of $\mathbf{m}$, tendency to rotate in the opposite sense.

7. $|\mathbf{v}| = |[0, 20, 0] \times [8, 6, 0]| = |[0, 0, -160]| = 160$

9. Zero volume in Fig. 191, which can happen in several ways.

11. $[0, 0, 7], \quad [0, 0, -7], \quad -4$

13. $[6, 2, 7], \quad [-6, -2, -7]$

15. $\mathbf{0}$

17. $[-32, -58, 34], \quad [-42, -63, 19]$

19. $1, \quad -1$

21. $[-48, -72, -168], \quad 12\sqrt{248} = 189.0, 189.0$

23. $0, \quad 0, \quad 13$

25. $\mathbf{m} = [-2, -2, 0] \times [2, 3, 0] = [0, 0, -10]$, $m = 10$ clockwise

27. $[6, 2, 0] \times [1, 2, 0] = [0, 0, 10]$

29. $\frac{1}{2}|[-12, 2, 6]| = \sqrt{46}$

31. $3x + 2y - z = 5$

33. $474/6 = 79$

Problem Set 9.4, page 380

1. Hyperbolas

3. Parallel straight lines (planes in space) $y = \frac{3}{4}x + c$

5. Circles, centers on the y-axis

7. Ellipses

9. Parallel planes

11. Elliptic cylinders

13. Paraboloids

Problem Set 9.5, page 390

1. Circle, center (3, 0), radius 2

3. Cubic parabola $x = 0, z = y^3$

5. Ellipse

7. Helix

9. A "Lissajous curve"

11. $\mathbf{r} = [3 + \sqrt{13}\cos t, 2 + \sqrt{13}\sin t, 1]$

13. $\mathbf{r} = [2 + t, 1 + 2t, 3]$

15. $\mathbf{r} = [t, 4t - 1, 5t]$

17. $\mathbf{r} = [\sqrt{2}\cos t, \sin t, \sin t]$

19. $\mathbf{r} = [\cosh t, (\sqrt{3}/2)\sinh t, -2]$

21. Use $\sin(-\alpha) = -\sin\alpha$.

25. $\mathbf{u} = [-\sin t, 0, \cos t]$. At P, $\mathbf{r}' = [-8, 0, 6]$. $\mathbf{q}(w) = [6 - 8w, i, 8 + 6w]$.

27. $\mathbf{q}(w) = [2 + w, \frac{1}{2} - \frac{1}{4}w, 0]$

29. $\sqrt{\mathbf{r}' \bullet \mathbf{r}'} = \cosh t, l = \sinh l = 1.175$

31. $\sqrt{\mathbf{r}' \bullet \mathbf{r}'} = a, l = a\pi/2$

33. Start from $\mathbf{r}(t) = [t, f(t)]$.

35. $\mathbf{v} = \mathbf{r}' = [1, 2t, 0], \quad |\mathbf{v}| = \sqrt{1 + 4t^2}, \quad \mathbf{a} = [0, 2, 0]$

37. $\mathbf{v}(0) = (\omega + 1)R\mathbf{i}, \mathbf{a}(0) = -\omega^2 R\mathbf{j}$

39. $\mathbf{v} = [-\sin t - 2\sin 2t, \cos t - 2\cos 2t], \quad |\mathbf{v}|^2 = 5 - 4\cos 3t,$

$$\mathbf{a} = [-\cos t - 4\cos 2t, -\sin t + 4\sin 2t], \text{ and } \mathbf{a}_{\tan} = \frac{6\sin 3t}{5 - 4\cos 3t}\mathbf{v}.$$

41. $\mathbf{v} = [-\sin t, 2\cos 2t, -2\sin 2t], \quad |\mathbf{v}|^2 = 4 + \sin^2 t,$

$$\mathbf{a} = [-\cos t, -4\sin 2t, -4\cos 2t], \text{ and } \mathbf{a}_{\tan} = \frac{\frac{1}{2}\sin 2t}{4 + \sin^2 t}\mathbf{v}.$$

43. 1 year $= 365 \cdot 86{,}400$ sec, $\quad R = 30 \cdot 365 \cdot 86{,}400/2\pi = 151 \cdot 10^6$ [km], $|\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R = 5.98 \cdot 10^{-6}$ [km/sec^2]

45. $R = 3960 + 80$ mi $= 2.133 \cdot 10^7$ ft, $\quad g = |\mathbf{a}| = \omega^2 R = |\mathbf{v}|^2/R, \quad |\mathbf{v}| = \sqrt{gR} = \sqrt{6.61 \cdot 10^8} = 25{,}700$ [ft/sec] $= 17{,}500$ [mph]

49. $\mathbf{r}(t) = [t, y(t), 0], \quad \mathbf{r}' = [1, y', 0]$ $\mathbf{r} \bullet \mathbf{r}' = 1 + y'^2$, etc.

51. $\dfrac{d\mathbf{r}}{ds} = \dfrac{d\mathbf{r}}{dt}\Big/\dfrac{ds}{dt}, \quad \dfrac{d^2\mathbf{r}}{ds^2} = \dfrac{d^2\mathbf{r}}{dt^2}\Big/\left(\dfrac{ds}{dt}\right)^2 + \cdots, \quad \dfrac{d^3\mathbf{r}}{ds^3} = \dfrac{d^3\mathbf{r}}{dt^3}\Big/\left(\dfrac{ds}{dt}\right)^3 + \cdots$

53. $3/(1 + 9t^2 + 9t^4)$

Problem Set 9.7, page 402

1. $[2y - 1, 2x + 2]$

3. $[-y/x^2, 1/x]$

5. $[4x^3, 4y^3]$

7. Use the chain rule.

9. Apply the quotient rule to each component and collect terms.

11. $[y, x]$, $[5, -4]$

13. $[2x/(x^2 + y^2), 2y/(x^2 + y^2)]$, $[0.16, 0.12]$

15. $[8x, 18y, 2z]$, $[40, -18, -22]$

17. For P on the x- and y-axes.

19. $[-1.25, 0]$

21. $[0, -e]$

23. Points with $y = 0, \pm\pi, \pm 2\pi, \cdots$.

25. $-\nabla T(P) = [0, 4, -1]$

31. $\nabla f = [32x, -2y]$, $\nabla f(P) = [160, -2]$

33. $[12x, 4y, 2z]$, $[60, 20, 10]$

35. $[-2x, -2y, 1]$, $[-6, -8, 1]$

37. $[2, 1] \bullet [1, -1]/\sqrt{5} = 1/\sqrt{5}$

39. $[1, 1, 1] \bullet [-3/125, 0, -4/125]/\sqrt{3} = -7/(125\sqrt{3})$

41. $\sqrt{8/3}$

43. $f = xyz$

45. $f = \int v_1\, dx + \int v_2\, dy + \int v_3\, dz$

Problem Set 9.8, page 405

1. $2x + 8y + 18z$; 7

3. 0, after simplification; solenoidal

5. $9x^2y^2z^2$; 1296

7. $-2e^x (\cos y)z$

9. (b) $(fv_1)_x + (fv_2)_y + (fv_3)_z = f[(v_1)_x + (v_2)_y + (v_3)_z] + f_x v_1 + f_y v_2 + f_z v_3$, etc.

11. $[v_1, v_2, v_3] = \mathbf{r}' = [x', y', z'] = [y, 0, 0]$, $z' = 0, z = c_3$, $y' = 0, y = c_2$, and $x' = y = c_2, x = c_2 t + c_1$. Hence as t increases from 0 to 1, this "shear flow" transforms the cube into a parallelepiped of volume 1.

13. div $(\mathbf{w} \times \mathbf{r}) = 0$ because v_1, v_2, v_3 do not depend on x, y, z, respectively.

15. $-2\cos 2x + 2\cos 2y$

17. 0

19. $2/(x^2 + y^2 + z^2)^2$

Problem Set 9.9, page 408

3. Use the definitions and direct calculation.

5. $[x(z^2 - y^2), y(x^2 - z^2), z(y^2 - x^2)]$

7. $e^{-x}[\cos y, \sin y, 0]$

9. curl $\mathbf{v} = [-6z, 0, 0]$ incompressible, $\mathbf{v} = \mathbf{r}' = [x', y', z'] = [0, 3z^2, 0]$, $x = c_1$, $z = c_3$, $y' = 3z^2 = 3c_3^2$, $y = 3c_3^2 t + c_2$

11. curl $\mathbf{v} = [0, 0, -3]$, incompressible, $x' = y$, $y' = -2x$, $2xx' + yy' = 0$, $x^2 + \frac{1}{2}y^2 = c, z = c_3$

13. curl $\mathbf{v} = 0$, irrotational, div $\mathbf{v} = 1$, compressible, $\mathbf{r} = [c_1 e^t, c_2 e^t, c_3 e^{-t}]$. Sketch it.

15. $[-1, -1, -1]$, same (why?)

17. $-yz - zx - xy, 0$ (why?), $-y - z - x$

19. $[-2z - y, -2x - z, -2y - x]$, same (why?)

Chapter 9 Review Questions and Problems, page 409

11. $-10, \quad 1080, \quad 1080, \quad 65$
13. $[-10, -30, 0], \quad [10, 30, 0], \quad \mathbf{0}, \quad 40$
15. $[-1260, -1830, -300], \quad [-210, 120, -540]$, undefined
17. $-125, \quad 125, \quad -125$
19. $[70, -40, -50], \quad 0, \quad \sqrt{35^2 + 20^2 + 25^2} = \sqrt{2250}$
21. $[-2, -6, -13]$
23. $\gamma_1 = \arccos(-10/\sqrt{65 \cdot 40}) = 1.7682 = -101.3°, \gamma_2 = 23.7°$
25. $[5, 2, 0] \cdot [4 - 1, 3 - 1, 0] = 19$
27. $\mathbf{v} \cdot \mathbf{w}/|\mathbf{w}| = 22/\sqrt{8} = 7.78$
29. $[0, 0, -14]$, tendency of clockwise rotation
31. 4
33. $1, \quad -2y$
35. 0, same (why?), $2(y^2 + x^2 - xz)$
37. $[0, -2, 0]$
39. $9/\sqrt{225} = \frac{3}{5}$

Problem Set 10.1, page 418

3. 4
5. $\mathbf{r} = [2 \cos t, \quad 2 \sin t], \quad 0 \leqq t \leqq \pi/2; \quad \frac{8}{5}$
7. "Exponential helix," $(e^{6\pi} - 1)/3$
9. $23.5, \quad 0$
11. $2e^{-t} + 2te^{-t^2}, \quad -2e^{-2} - e^{-4} + 3$
15. $18\pi, \quad \frac{4}{3}(4\pi)^3, \quad 18\pi$
17. $[4 \cos t, \quad + \sin t, \quad \sin t, \quad 4 \cos t], \quad [2, 2, 0]$
19. $144t^4, \quad 1843.2$

Problem Set 10.2, page 425

3. $\sin \frac{1}{2}x \cos 2y, \quad 1 - 1/\sqrt{2} = 0.293$
5. $e^{xy} \sin z, \quad e - 0$
7. $\cosh 1 - 2 = -0.457$
9. $e^x \cosh y + e^z \sinh y, \quad e - (\cosh 1 + \sinh 1) = 0$
13. $e^{a^2} \cos 2b$
15. Dependent, $x^2 \neq -4y^2$, etc.
17. Dependent, $4 \neq 0$, etc.
19. $\sin(a^2 + 2b^2 + c^2)$

Problem Set 10.3, page 432

3. $8y^3/3, \quad 54$
5. $\int_0^1 [x - x^3 - (x^2 - x^5)]\, dx = \frac{1}{12}$
7. $\cosh 2x - \cosh x, \quad \frac{1}{2} \sinh 4 - \sinh 2$
9. $36 + 27y^2, \quad 144$
11. $z = 1 - r^2, \quad dx\, dy = r\, dr\, d\theta$, *Answer*: $\pi/2$
13. $\bar{x} = 2b/3, \quad \bar{y} = h/3$
15. $\bar{x} = 0, \quad \bar{y} = 4r/3\pi$
17. $I_x = bh^3/12, \quad I_y = b^3h/4$
19. $I_x = (a + b)h^3/24, \quad I_y = h(a^4 - b^4)/(48(a - b))$

Problem Set 10.4, page 438

1. $(-1 - 1) \cdot \pi/4 = -\pi/2$
3. $9(e^2 - 1) - \frac{8}{3}(e^3 - 1)$
5. $2x - 2y, \quad 2x(1 - x^2) - (2 - x^2)^2 + 1, \quad x = -1 \cdots 1, \quad -\frac{56}{15}$
7. 0. Why?
9. $\frac{16}{5}$
13. $\nabla^2 w = \cosh x, \quad y = x/2 \cdots 2, \quad \frac{1}{2} \cosh 4 - \frac{1}{2}$

15. $\nabla^2 w = 6xy, \quad 3x(10 - x^2)^2 - 3x, \quad 486$
17. $\nabla^2 w = 6x - 6y, \quad -38.4$
19. $|\text{grad } w|^2 = e^{2x}, \quad \frac{5}{2}(e^4 - 1)$

Problem Set 10.5, page 442

1. Straight lines, $\mathbf{k}$
3. $z = c\sqrt{x^2 + y^2}$, circles, straight lines, $[-cu \cos v, \quad -cu \sin v, \quad u]$
5. $z = x^2 + y^2$, circles, parabolas, $[-2u^2 \cos v, \quad -2u^2 \sin v, \quad u]$
7. $x^2/a^2 + y^2/b^2 + z^2/c^2 = 1, \quad [bc \cos^2 v \cos u, \quad ac \cos^2 v \sin u, \quad ab \sin v \cos v]$, ellipses
11. $[\tilde{u}, \quad \tilde{v}, \quad \tilde{u}^2, \ + \tilde{v}^2], \quad \tilde{\mathbf{N}} = [-2\tilde{u}, \quad -2\tilde{v}, \quad 1]$
13. Set $x = u$ and $y = v$.
15. $[2 + 5 \cos u, \quad -1 + 5 \sin u, \quad v], \quad [5 \cos u, \quad 5 \sin u, \quad 0]$
17. $[a \cos v \cos u, \quad -2.8 + a \cos v \sin u, \quad 3.2 + a \sin v], \quad a = 1.5;$ $[a^2 \cos^2 v \cos u, \quad a^2 \cos^2 v \sin u, \quad a^2 \cos v \sin v]$
19. $[\cosh u, \quad \sinh u, \quad v], \quad [\cosh u, \quad -\sinh u, \quad 0]$

Problem Set 10.6, page 450

1. $\mathbf{F}(\mathbf{r}) \bullet \mathbf{N} = [-u^2, \quad v^2, \quad 0] \bullet [-3, \quad 2, \quad 1] = 3u^2 + 2v^2, \quad 29.5$
3. $\mathbf{F}(\mathbf{r}) \bullet \mathbf{N} = \cos^3 v \cos u \sin u$ from (3), Sec. 10.5. *Answer*: $\frac{1}{3}$
5. $\mathbf{F}(\mathbf{r}) \bullet \mathbf{N} = -u^3, -128\pi$
7. $\mathbf{F} \bullet \mathbf{N} = [0, \quad \sin u, \quad \cos v] \bullet [1, -2u, 0], \quad 4 + (-2 + \pi^2/16 - \pi/2)\sqrt{2} = -0.1775$
9. $\mathbf{r} = [2 \cos u, \quad 2 \sin u, \quad v], \quad 0 \leqq u \leqq \pi/4, \quad 0 \leqq v \leqq 5$. Integrate $2 \sinh v \sin u$ to get $2(1 - 1/\sqrt{2})(\cosh 5 - 1) = 42.885$.
13. $7\pi^3/\sqrt{6} = 88.6$
15. $G(\mathbf{r}) = (1 + 9u^4)^{3/2}, \quad |\mathbf{N}| = (1 + 9u^4)^{1/2}$. *Answer*: 54.4
21. $I_{x=y} = \iint_S [\frac{1}{2}(x - y)^2 + z^2]\, \sigma\, dA$
23. $[u \cos v, \quad u \sin v, \quad u], \quad \int_0^{2\pi}\int_0^h u^2 \cdot u\sqrt{2}\, du\, dv = \frac{\pi}{\sqrt{2}} h^4$
25. $[\cos u \cos v, \quad \cos u \sin v, \quad \sin u], \quad dA = (\cos u)\, du\, dv$, B the z-axis, $\quad I_B = 8\pi/3$, $I_K = I_B + 1^2 \cdot 4\pi = 20.9$.

Problem Set 10.7, page 457

1. 224
3. $-e^{-1-z} + e^{-y-z}, \quad -2e^{-1-z} + e^{-z}, \quad 2e^{-3} - e^{-2} - 2e^{-1} + 1$
5. $\frac{1}{2}(\sin 2x)(1 - \cos 2x), \quad \frac{1}{8}, \quad \frac{3}{4}$
7. $[r \cos u \cos v, \quad \cos u \sin v, \quad r \sin u], \quad dV = r^2 \cos u\, dr\, du\, dv, \quad \sigma = v, \quad 2\pi^2 a^3/3$
9. div $\mathbf{F} = 2x + 2z$, 48
11. $12(e - 1/e) = 24 \sinh 1$
13. div $\mathbf{F} = -\sin z$, 0
15. $1/\pi + \frac{5}{24} = 0.5266$
17. $h^4\pi/2$
19. $8abc(b^2 + c^2)/3$
21. $(a^4/4) \cdot 2\pi \cdot h = ha^4\pi/2$
23. $h^5\pi/10$
25. Do Prob. 20 as the last one.

Problem Set 10.8, page 462

1. $x = 0, y = 0, z = 0$, no contributions. $x = a$: $\partial f/\partial n = \partial f/\partial x = -2x = -2a$, etc. Integrals $x = a$: $(-2a)bc$, $y = b$: $(-2b)ac$, $z = c$: $(4c)\ ab$. Sum 0

3. The volume integral of $8y^2 + [0, 8y] \cdot [2x, 0] = 8y^2$ is $8y^3/3 = \frac{8}{3}$. The surface integral of $f\partial g/\partial n = f \cdot 2x = 2f = 8y^2$ over $x = 1$ is $8y^3/3 = \frac{8}{3}$. Others 0.

5. The volume integral of $6y^2 \cdot 4 - 2x^2 \cdot 12$ is 0; $8(x = 1)$, $-8(y = 1)$, others 0.

7. $\mathbf{F} = [x, \ 0, \ 0]$, $\operatorname{div} \mathbf{F} = 1$, use (2*), Sec. 10.7, etc.

9. $z = 0$ and $z = \sqrt{a^2 - x^2 - y^2} = \sqrt{a^2 - r^2}$, $dx\,dy = r\,dr\,d\theta$, $-2\pi \cdot \frac{1}{2}(a^2 - r^2)^{3/2} \cdot \frac{2}{3}\big|_0^a = \frac{2}{3}\pi a^3$

11. $r = a$, $\phi = 0$, $\cos\phi = 1$, $v = \frac{1}{3}a \cdot (4\pi a^2)$

Problem Set 10.9, page 468

1. S: $z = y\ (0 \leqq x \leqq 1, 0 \leqq y \leqq 4)$, $[0, 2z, -2z] \cdot [0, -1, 1]$, ± 20

3. $[2e^{-z}\cos y, \ -e^{-z}, \ 0] \cdot [0, \ -y, \ 1] = ye^{-z}$, $\pm(2 - 2/\sqrt{e})$

5. $[0, 2z, \frac{3}{2}] \cdot [0, 0, 1] = \frac{3}{2}$, $\pm\frac{3}{2}a^2$

7. $[-e^z, \ -e^x, \ -e^y] \cdot [-2x, \ 0, \ 1]$, $\pm(e^4 - 2e + 1)$

9. The sides contribute a, $3a^2/2$, $-a$, 0.

11. -2π; curl $\mathbf{F} = \mathbf{0}$

13. $5\mathbf{k}$, 80π

15. $[0, -1, 2x - 2y] \cdot [0, 0, 1]$, $\frac{1}{3}$

17. $\mathbf{r} = [\cos u, \ \sin u, \ v]$, $[-3v^2, \ 0, \ 0] \cdot [\cos u, \ \sin u, \ 0]$, -1

19. $\mathbf{r} = [u\cos v, \ u\sin v, \ u]$, $0 \leqq u \leqq 1$, $0 \leqq v \leqq \pi/2$, $[-e^z, \ 1, \ 0] \cdot [-u\cos v, \ -u\sin v, \ u]$. *Answer*: 1/2

Chapter 10 Review Questions and Problems, page 469

11. $\mathbf{r} = [4 - 10t, \ 2 + 8t]$, $\mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = [2(4 - 10t)^2, \ -4(2t + 8t)^2] \cdot [-10, \ 8]\,dt$; $-4528/3$. Or using exactness.

13. Not exact, curl $\mathbf{F} = (5\cos x)\mathbf{k}$, ± 10

15. 0 since curl $\mathbf{F} = \mathbf{0}$

17. By Stokes, $\pm 18\pi$

19. $\mathbf{F} = \operatorname{grad}(y^2 + xz)$, 2π

21. $M = 8$, $\bar{x} = \frac{8}{5}$, $\bar{y} = \frac{16}{5}$

23. $M = \frac{63}{20}$, $\bar{x} = \frac{8}{7} = 1.14$, $\bar{y} = \frac{118}{49} = 2.41$

25. $M = 4k/15$, $\bar{x} = \frac{5}{16}$, $\bar{y} = \frac{4}{7}$

27. $288(a + b + c)\pi$

29. div $\mathbf{F} = 20 + 6z^2$. *Answer*: 21

31. $24 \sinh 1 = 28.205$

33. Direct integration, $\frac{224}{3}$

35. 72π

Problem Set 11.1, page 482

1. $2\pi, 2\pi, \pi, \pi, 1, 1, \frac{1}{2}, \frac{1}{2}$

5. There is no *smallest* $p > 0$.

13. $\frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right) + 2\left(\sin x + \frac{1}{3}\sin 3x + \frac{1}{5}\sin 5x + \cdots\right)$

15. $\frac{4}{3}\pi^2 + 4\left(\cos x + \frac{1}{4}\cos 2x + \frac{1}{9}\cos 3x + \cdots\right) - 4\pi\left(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \cdots\right)$

17. $\frac{\pi}{2} + \frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$

19. $\frac{\pi}{4} - \frac{2}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right) + \sin x - \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x - + \cdots$

21. $2(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \frac{1}{4}\sin 4x + \frac{1}{5}\sin 5x + \cdots)$

Problem Set 11.2, page 490

1. Neither, even, odd, odd, neither **3.** Even **5.** Even

9. Odd, $L = 2$, $\frac{4}{\pi}\left(\sin\frac{\pi x}{2} + \frac{1}{3}\sin\frac{3\pi x}{2} + \frac{1}{5}\sin\frac{5\pi x}{2} + \cdots\right)$

11. Even, $L = 1$, $\frac{1}{3} - \frac{4}{\pi^2}\left(\cos \pi x - \frac{1}{4}\cos 2\pi x + \frac{1}{9}\cos 3\pi x - + \cdots\right)$

13. Rectifier, $L = \frac{1}{2}$, $\frac{1}{8} - \frac{1}{\pi^2}\left(\cos 2\pi x + \frac{1}{9}\cos 6\pi x + \frac{1}{25}\cos 10\pi x + \cdots\right) + \frac{1}{\pi}\left(\frac{1}{2}\sin 2\pi x - \frac{1}{4}\sin 4\pi x + \frac{1}{6}\sin 6\pi x - \frac{1}{8}\sin 8\pi x + - \cdots\right)$

15. Odd, $L = \pi$, $\frac{4}{\pi}\left(\sin x - \frac{1}{9}\sin 3x + \frac{1}{25}\sin 5x - + \cdots\right)$

17. Even, $L = 1$, $\frac{1}{2} + \frac{4}{\pi^2}\left(\cos \pi x + \frac{1}{9}\cos 3\pi x + \frac{1}{25}\cos 5\pi x + \cdots\right)$

19. $\frac{3}{8} + \frac{1}{2}\cos 2x + \frac{1}{8}\cos 4x$

23. $L = 4$, **(a)** 1, **(b)** $\frac{4}{\pi}\left(\sin\frac{\pi x}{4} + \frac{1}{3}\sin\frac{3\pi x}{4} + \frac{1}{5}\sin\frac{5\pi x}{4} + \cdots\right)$

25. $L = \pi$, **(a)** $\frac{\pi}{2} + \frac{4}{\pi}\left(\cos x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x + \cdots\right)$,
(b) $2(\sin x + \frac{1}{2}\sin 2x + \frac{1}{3}\sin 3x + \frac{1}{4}\sin 4x + \cdots)$

27. $L = \pi$, **(a)** $\frac{3\pi}{8} + \frac{2}{\pi}\left(\cos x - \frac{1}{2}\cos 2x + \frac{1}{9}\cos 3x + \frac{1}{25}\cos 5x - \frac{1}{18}\cos 6x + \frac{1}{49}\cos 7x + \frac{1}{81}\cos 9x - \frac{1}{50}\cos 10x + \frac{1}{121}\cos 11x + \cdots\right)$
(b) $\left(1 + \frac{2}{\pi}\right)\sin x + \frac{1}{2}\sin 2x + \left(\frac{1}{3} - \frac{2}{9\pi}\right)\sin 3x + \frac{1}{4}\sin 4x + \left(\frac{1}{5} + \frac{2}{25\pi}\right)\sin 5x + \frac{1}{6}\sin 6x + \cdots$

29. Rectifier, $L = \pi$,
(a) $\frac{2}{\pi} - \frac{4}{\pi}\left(\frac{1}{1\cdot 3}\cos x + \frac{1}{3\cdot 5}\cos 3x + \frac{1}{5\cdot 7}\cos 5x + \cdots\right)$, **(b)** $\sin x$

Problem Set 11.3, page 494

3. The output becomes a pure cosine series.

5. For A_n this is similar to Fig. 54 in Sec. 2.8, whereas for the phase shift B_n the sense is the same for all n.

7. $y = C_1 \cos \omega t + C_2 \sin \omega t + a(\omega) \sin t$, $a(\omega) = 1/(\omega^2 - 1) = -1.33$, $-5.26, 4.76, 0.8, 0.01$. Note the change of sign.

11. $y = C_1 \cos \omega t + C_2 \sin \omega t + \frac{4}{\pi}\left(\frac{1}{\omega^2 - 9} \sin t + \frac{1}{\omega^2 - 49} \sin 3t + \frac{1}{\omega^2 - 121} \sin 5t + \cdots\right)$

13. $y = \sum_{n=1}^{N} (A_n \cos nt + B_n \sin nt)$, $A_n = [(1 - n^2)a_n - nb_n c]/D_n$, $B_n = [(1 - n^2)b_n + nca_n]/D_n$, $D_n = (1 - n^2)^2 + n^2c^2$

15. $b_n = (-1)^{n+1} \cdot 12/n^3$ (n odd), $y = \sum_{n=1}^{\infty} (A_n \cos nt + B_n \sin nt)$, $A_n = (-1)^n \cdot 12nc/n^3D_n$, $B_n = (-1)^{n+1} \cdot 12(1 - n^2)/(n^3D_n)$ with D_n as in Prob. 13.

17. $I = 50 + A_1 \cos t + B_1 \sin t + A_3 \cos 3t + B_3 \sin 3t + \cdots$, $A_n = (10 - n^2)a_n/D_n$, $B_n = 10na_n/D_n$, $a_n = -400/(n^2\pi)$, $D_n = (n^2 - 10)^2 + 100n^2$

19. $I(t) = \sum_{n=1}^{\infty} (A_n \cos nt + B_n \sin nt)$, $A_n = (-1)^{n+1} \frac{2400(10 - n^2)}{n^2D_n}$, $B_n = (-1)^{n+1} \frac{24{,}000}{nD_n}$, $D_n = (10 - n^2)^2 + 100n^2$

Section 11.4, page 498

3. $F = \frac{\pi}{2} - \frac{4}{\pi}\left(\cos x + \frac{1}{9} \cos 3x + \frac{1}{25} \cos 5x + \cdots\right)$, $E^* = 0.0748$, $0.0748, 0.0119, 0.0119, 0.0037$

5. $F = \frac{4}{\pi}\left(\sin x + \frac{1}{3} \sin 3x + \frac{1}{5} \sin 5x + \cdots\right)$, $E^* = 1.1902, 1.1902, 0.6243, 0.6243$, 0.4206 (0.1272 when $N = 20$)

7. $F = 2[(\pi^2 - 6) \sin x - \frac{1}{8}(4\pi^2 - 6) \sin 2x + \frac{1}{27}(9\pi^2 - 6) \sin 3x - + \cdots]$; $E^* = 674.8, 454.7, 336.4, 265.6, 219.0$. Why is E^* so large?

Section 11.5, page 503

3. Set $x = ct + k$.

5. $x = \cos \theta$, $dx = -\sin \theta \, d\theta$, etc.

7. $\lambda_m = (m\pi/10)^2$, $m = 1, 2, \cdots$; $y_m = \sin(m\pi x/10)$

9. $\lambda = [(2m + 1)\pi/(2L)]^2$, $m = 0, 1, \cdots$, $y_m = \sin((2m + 1)\pi x/(2L))$

11. $\lambda_m = m^2$, $m = 1, 2, \cdots$, $y_m = x \sin(m \ln |x|)$

13. $p = e^{8x}$, $q = 0$, $r = e^{8x}$, $\lambda_m = m^2$, $y_m = e^{-4x} \sin mx$, $m = 1, 2, \cdots$

Section 11.6, page 509

1. $8(P_1(x) - P_3(x) + P_5(x))$

3. $\frac{4}{5}P_0(x) - \frac{4}{7}P_2(x) - \frac{8}{35}P_4(x)$

9. $-0.4775P_1(x) - 0.6908P_3(x) + 1.844P_5(x) - 0.8236P_7(x) + 0.1658P_9(x) + \cdots$, $m_0 = 9$. ***Rounding*** seems to have considerable influence in Probs. 8–13.

11. $0.7854P_0(x) - 0.3540P_2(x) + 0.0830P_4(x) - \cdots, m_0 = 4$
13. $0.1212P_0(x) - 0.7955P_2(x) + 0.9600P_4(x) - 0.3360P_6(x) + \cdots, m_0 = 8$
15. (c) $a_m = (2/J_1^2(\alpha_{0,m}))(J_1(\alpha_{0,m})/\alpha_{0,m}) = 2/(\alpha_{0,m}J_1(\alpha_{0,m}))$

Section 11.7, page 517

1. $f(x) = \pi e^{-x} (x > 0)$ gives $A = \int_0^\infty e^{-v} \cos wv \, dv = \dfrac{1}{1 + w^2}$, $B = \dfrac{w}{1 + w^2}$ (see Example 3), etc.

3. Use (11); $B = \dfrac{2}{\pi} \int_0^\infty \dfrac{\pi}{2} \sin wv \, dv = \dfrac{1 - \cos \pi w}{w}$

5. $B(w) = \dfrac{2}{\pi} \int_0^1 \dfrac{1}{2} \pi v \sin wv \, dv = \dfrac{\sin w - w \cos w}{w^2}$

7. $\dfrac{2}{\pi} \int_0^\infty \dfrac{\sin w \cos xw}{w} dw$

9. $A(w) = \dfrac{2}{\pi} \int_0^\infty \dfrac{\cos wv}{1 + v^2} dv = e^{-w} \ (w > 0)$

11. $\dfrac{2}{\pi} \int_0^\infty \dfrac{\cos \pi w + 1}{1 - w^2} \cos xw \, dw$

15. For $n = 1, 2, 11, 12, 31, 32, 49, 50$ the value of $\text{Si}(n\pi) - \pi/2$ equals $0.28, -0.15, 0.029, -0.026, 0.0103, -0.0099, 0.0065, -0.0064$ (rounded).

17. $\dfrac{2}{\pi} \int_0^\infty \dfrac{1 - \cos w}{w} \sin xw \, dw$

19. $\dfrac{2}{\pi} \int_0^\infty \dfrac{w - e(w \cos w - \sin w)}{1 + w^2} \sin xw \, dw$

Section 11.8, page 522

1. $\hat{f}_c(w) = \sqrt{(2/\pi)} \, (2 \sin w - \sin 2w)/w$
3. $\hat{f}_c(w) = \sqrt{(2/\pi)} \, (\cos 2w + 2w \sin 2w - 1)/w^2$
5. $\hat{f}_c(w) = \sqrt{\dfrac{2}{\pi}} \dfrac{(w^2 - 2) \sin w + 2w \cos w}{w^3}$
7. Yes. No
9. $\sqrt{2/\pi} \, w/(a^2 + w^2)$
11. $\sqrt{2/\pi} \, ((2 - w^2) \cos w + 2w \sin w - 2)/w^3$
13. $\mathcal{F}_s(e^{-x}) = \dfrac{1}{w}\left(-\mathcal{F}_c(e^{-x}) + \sqrt{\dfrac{2}{\pi}} \cdot 1\right) = \dfrac{1}{w}\left(\sqrt{\dfrac{2}{\pi}} \cdot \dfrac{1}{w^2 + 1} + \sqrt{\dfrac{2}{\pi}}\right) = \sqrt{\dfrac{2}{\pi}} \dfrac{w}{w^2 + 1}$

Problem Set 11.9, page 533

3. $i(e^{-ibw} - e^{-iaw})/(w\sqrt{2\pi})$ if $a < b$; 0 otherwise
5. $[e^{(1-iw)a} - e^{-(1-iw)a}]/(\sqrt{2\pi}(1 - iw))$
7. $(e^{-iaw}(1 + iaw) - 1)/(\sqrt{2\pi} w^2)$
9. $\sqrt{2/\pi}(\cos w + w \sin w - 1)/w^2$
11. $i\sqrt{2/\pi} \, (\cos w - 1)/w$
13. $e^{-w^2/2}$ by formula 9

17. No, the assumptions in Theorem 3 are not satisfied.

19. $[f_1 + f_2 + f_3 + f_4, \quad f_1 - if_2 - f_3 + if_4, \quad f_1 - f_2 + f_3 - f_4, \quad f_1 + if_2 - f_3 - if_4]$

21. $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} f_1 \\ f_2 \end{bmatrix} = \begin{bmatrix} f_1 + f_2 \\ f_1 - f_2 \end{bmatrix}$

Chapter 11 Review Questions and Problems, page 537

11. $1 + \dfrac{4}{\pi}\left(\sin\dfrac{\pi x}{2} + \dfrac{1}{3}\sin\dfrac{3\pi x}{2} + \dfrac{1}{5}\sin\dfrac{5\pi x}{2} + \cdots\right)$

13. $\dfrac{1}{4} - \dfrac{2}{\pi^2}\left(\cos \pi x + \dfrac{1}{9}\cos 3\pi x + \dfrac{1}{25}\cos 5\pi x + \cdots\right) +$
$\dfrac{1}{\pi}\left(\sin \pi x - \dfrac{1}{2}\sin 2\pi x + \dfrac{1}{3}\sin 3\pi x - + \cdots\right)$

15. $\cosh x$, $\sinh x$ $(-5 < x < 5)$, respectively

17. Cf. Sec. 11.1.

19. $\dfrac{1}{2} - \dfrac{4}{\pi^2}\left(\cos \pi x + \dfrac{1}{9}\cos 3\pi x + \cdots\right), \quad \dfrac{2}{\pi}\left(\sin \pi x - \dfrac{1}{2}\sin 2\pi x + - \cdots\right)$

21. $y = C_1 \cos \omega t + C_2 \sin \omega t + \dfrac{\pi^2}{\omega^2} - 12\left(\dfrac{\cos t}{\omega^2 - 1} - \dfrac{1}{4}\cdot\dfrac{\cos 2t}{\omega^2 - 4} + \dfrac{1}{9}\cdot\dfrac{\cos 3t}{\omega^2 - 9}\right.$
$\left. - \dfrac{1}{16}\cdot\dfrac{\cos 4t}{\omega^2 - 16} + - \cdots\right)$

23. 0.82, 0.50, 0.36, 0.28, 0.23

25. 0.0076, 0.0076, 0.0012, 0.0012, 0.0004

27. $\dfrac{1}{\pi}\displaystyle\int_0^\infty \dfrac{(\cos w + w \sin w - 1)\cos wx + (\sin w - w \cos w)\sin wx}{w^2}\, dw$

29. $\sqrt{2/\pi}\,(\cos aw - \cos w + aw \sin aw - w \sin w)/w^2$

Problem Set 12.1, page 542

1. $L(c_1u_1 + c_2u_2) = c_1L(u_1) + c_2L(u_2) = c_1 \cdot 0 + c_2 \cdot 0 = 0$

3. $c = 2$

5. $c = a/b$

7. Any c and ω

9. $c = \pi/25$

15. $u = 110 - (110/\ln 100)\ln(x^2 + y^2)$

17. $u = a(y)\cos 4\pi x + b(y)\sin 4\pi x$

19. $u = c(x)\, e^{-y^3/3}$

21. $u = e^{-3y}(a(x)\cos 2y + b(x)\sin 2y) + 0.1e^{3y}$

23. $u = c_1(y)x + c_2(y)/x^2$ (Euler–Cauchy)

25. $u(x, y) = axy + bx + cy + k$; a, b, c, k arbitrary constants

Problem Set 12.3, page 551

5. $k \cos 3\pi t \sin 3\pi x$

7. $\dfrac{8k}{\pi^3}\left(\cos \pi t \sin \pi x + \dfrac{1}{27}\cos 3\pi t \sin 3\pi x + \dfrac{1}{125}\cos 5\pi t \sin 5\pi x + \cdots\right)$

9. $\dfrac{0.8}{\pi^2}\left(\cos \pi t \sin \pi x - \dfrac{1}{9}\cos 3\pi t \sin 3\pi x + \dfrac{1}{25}\cos 5\pi t \sin 5\pi x - + \cdots\right)$

11. $\dfrac{2}{\pi^2}\left((2-\sqrt{2})\cos \pi t \sin \pi x - \dfrac{1}{9}(2+\sqrt{2})\cos 3\pi t \sin 3\pi x\right.$
$\left. + \dfrac{1}{25}(2+\sqrt{2})\cos 5\pi t \sin 5\pi x - + \cdots\right)$

13. $\dfrac{4}{\pi^3}\left((4-\pi)\cos \pi t \sin \pi x + \cos 2\pi t \sin 2\pi x + \dfrac{4+3\pi}{27}\cos 3\pi t \sin 3\pi x\right.$
$\left. + \dfrac{4-5\pi}{125}\cos 5\pi t \sin 5\pi x + \cdots\right)$. No terms with $n = 4, 8, 12, \cdots$.

17. $u = \dfrac{8L^2}{\pi^3}\left(\cos\left[c\left(\dfrac{\pi}{L}\right)^2 t\right]\sin\dfrac{\pi x}{L} + \dfrac{1}{3^3}\cos\left[c\left(\dfrac{3\pi}{L}\right)^2 t\right]\sin\dfrac{3\pi x}{L} + \cdots\right)$

19. **(a)** $u(0, t) = 0$, **(b)** $u(L, t) = 0$, **(c)** $u_x(0, t) = 0$, **(d)** $u_x(L, t) = 0$. $C = -A$, $D = -B$ from (a), (c). Insert this. The coefficient determinant resulting from (b), (d) must be zero to have a nontrivial solution. This gives (22).

Problem Set 12.4, page 556

3. $c^2 = 300/[0.9/(2 \cdot 9.80)] = 80.83^2$ [m^2/sec^2]
9. Elliptic, $u = f_1(y + 2ix) + f_2(y - 2ix)$
11. Parabolic, $u = xf_1(x - y) + f_2(x - y)$
13. Hyperbolic, $u = f_1(y - 4x) + f_2(y - x)$
15. Hyperbolic, $xy'^2 + yy' = 0$, $y = v$, $xy = w$, $u_w = z$, $u = \dfrac{1}{y}f_1(xy) + f_2(y)$
17. Elliptic, $u = f_1(y - (2 - i)x) + f_2(y - (2 + i)x)$. Real or imaginary parts of any function u of this form are solutions. Why?

Problem Set 12.6, page 566

3. $u_1 = \sin x\, e^{-t}$, $u_2 = \sin 2x\, e^{-4t}$, $u_3 = \sin 3x\, e^{-9t}$ differ in rapidity of decay.
5. $u = \sin 0.1\pi x\, e^{-1.752\pi^2 t/100}$
7. $u = \dfrac{800}{\pi^3}\left(\sin 0.1\pi x\, e^{-0.01752\pi^2 t} + \dfrac{1}{3^3}\sin 0.3\pi x\, e^{-0.01752(3\pi)^2 t} + \cdots\right)$
9. $u = u_{\text{I}} + u_{\text{II}}$, where $u_{\text{II}} = u - u_{\text{I}}$ satisfies the boundary conditions of the text, so that $u_{\text{II}} = \displaystyle\sum_{n=1}^{\infty} B_n \sin\frac{n\pi x}{L}\, e^{-(cn\pi/L)^2 t}$, $B_n = \dfrac{2}{L}\displaystyle\int_0^L [f(x) - u_{\text{I}}(x)]\sin\frac{n\pi x}{L}\,dx$.
11. $F = A\cos px + B\sin px$, $F'(0) = Bp = 0$, $B = 0$, $F'(L) = -Ap\sin pL = 0$, $p = n\pi/L$, etc.
13. $u = 1$
15. $\dfrac{1}{2} + \dfrac{4}{\pi^2}\left(\cos x\, e^{-t} + \dfrac{1}{9}\cos 3x\, e^{-9t} + \dfrac{1}{25}\cos 5x\, e^{-25t} + \cdots\right)$
17. $-\dfrac{K\pi}{L}\displaystyle\sum_{n=1}^{\infty} nB_n e^{-\lambda_n^2 t}$
19. $u = 1000\,(\sin\frac{1}{2}\pi x \sinh\frac{1}{2}\pi y)/\sinh \pi$
21. $u = \dfrac{100}{\pi}\displaystyle\sum_{n=1}^{\infty}\frac{1}{(2n-1)\sinh(2n-1)\pi}\sin\frac{(2n-1)\pi x}{24}\sinh\frac{(2n-1)\pi y}{24}$

23. $u = A_0 x + \sum_{n=1}^{\infty} A_n \frac{\sinh (n\pi x/24)}{\sinh n\pi} \cos \frac{n\pi y}{24}$,

$$A_0 = \frac{1}{24^2}\int_0^{24} f(y)\, dy, \quad A_n = \frac{1}{12}\int_0^{24} f(y) \cos \frac{n\pi y}{24}\, dy$$

25. $\sum_{n=1}^{\infty} A_n \sin \frac{n\pi x}{a} \sinh \frac{n\pi(b-y)}{a}$, $A_n = \frac{2}{a \sinh (n\pi b/a)} \int_0^a f(x) \sin \frac{n\pi x}{a}\, dx$

Problem Set 12.7, page 574

3. $A = \frac{2}{\pi}\int_0^{\infty} \frac{\cos pv}{1+v^2}\, dv = \frac{2}{\pi} \cdot \frac{\pi}{2} e^{-p}$, $u = \int_0^{\infty} e^{-p-c^2p^2t} \cos px\, dp$

5. $A = \frac{2}{\pi}\int_0^1 v \cos pv\, dv = \frac{2}{\pi} \cdot \frac{\cos p + p \sin p - 1}{p^2}$, etc.

7. $A = \frac{2}{\pi}\int_0^{\infty} \frac{\sin v}{v} \cos pv\, dv = \frac{2}{\pi} \cdot \frac{\pi}{2} = 1$ if $0 < p < 1$ and 0 if $p > 1$,

$$u = \int_0^1 \cos px\, e^{-c^2p^2t}\, dp$$

9. Set $w = -v$ in (21) to get erf $(-x) = -$erf x.

13. In (12) the argument $x + 2cz\sqrt{t}$ is 0 (the point where f jumps) when $z = -x/(2c\sqrt{t})$. This gives the lower limit of integration.

15. Set $w = s/\sqrt{2}$ in (21).

Problem Set 12.9, page 584

1. **(a)**, **(b)** It is multiplied by $\sqrt{2}$. **(c)** Half

5. $B_{mn} = (-1)^{n+1} 8/(mn\pi^2)$ if m odd, 0 if m even

7. $B_{mn} = (-1)^{m+n} 4ab/(mn\pi^2)$

11. $u = 0.1 \cos \sqrt{20}t \sin 2x \sin 4y$

13. $\frac{6.4}{\pi^2} \sum_{\substack{m=1 \\ m,n}}^{\infty} \sum_{\substack{n=1 \\ \text{odd}}}^{\infty} \frac{1}{m^3n^3} \cos (t\sqrt{m^2+n^2}) \sin mx \sin ny$

17. $c\pi\sqrt{260}$ (corresponding eigenfunctions $F_{4,16}$ and $F_{16,14}$), etc.

19. $\cos\left(\pi t\sqrt{\frac{36}{a^2} + \frac{4}{b^2}}\right) \sin \frac{6\pi x}{a} \sin \frac{4\pi y}{b}$

Problem Set 12.10, page 591

5. $110 + \frac{440}{\pi}\left(r \cos \theta - \frac{1}{3} r^3 \cos 3\theta + \frac{1}{5} r^5 \cos 5\theta - + \cdots\right)$

7. $55\pi - \frac{440}{\pi}\left(r \cos \theta + \frac{1}{9} r^3 \cos 3\theta + \frac{1}{25} r^5 \cos 5\theta + \cdots\right)$

11. Solve the problem in the disk $r < a$ subject to u_0 (given) on the upper semicircle and $-u_0$ on the lower semicircle.

$$u = \frac{4u_0}{\pi}\left(\frac{r}{a}\sin\theta + \frac{1}{3a^3}r^3\sin 3\theta + \frac{1}{5a^5}r^5\sin 5\theta + \cdots\right)$$

13. Increase by a factor $\sqrt{2}$

15. $T = 6.826\rho R^2 f_1^2$

17. No

25. $\alpha_{11}/(2\pi) = 0.6098$; See Table A1 in App. 5.

Problem Set 12.11, page 598

5. $A_4 = A_6 = A_8 = A_{10} = 0, \quad A_5 = 605/16, \quad A_7 = -4125/128, \quad A_9 = 7315/256$

9. $\nabla^2 u = u'' + 2u'/r = 0, \quad u''/u' = -2/r, \quad \ln|u'| = -2\ln|r| + c_1,$
$u' = \tilde{c}/r^2, u = c/r + k$

13. $u = 320/r + 60$ is smaller than the potential in Prob. 12 for $2 < r < 4$.

17. $u = 1$

19. $\cos 2\phi = 2\cos^2\phi - 1, 2w^2 - 1 = \frac{4}{3}P_2(w) - \frac{1}{3}, u = \frac{4}{3}r^2P_2(\cos\phi) - \frac{1}{3}$

25. Set $1/r = \rho$. Then $u(\rho, \theta, \phi) = rv(r, \theta, \phi), u_\rho = (v + rv_r)(-1/\rho^2),$
$u_{\rho\rho} = (2v_r + rv_{rr})(1/\rho^4) + (v + rv_r)(2/\rho^3), u_{\rho\rho} + (2/\rho)u_\rho = r^5(v_{rr} + (2/r)v_r).$
Substitute this and $u_{\phi\phi} = rv_{\phi\phi}$ etc. into (7) [written in terms of ρ] and divide by r^5.

Problem Set 12.12, page 602

5. $W = \dfrac{c(s)}{x^s} + \dfrac{x}{s^2(s+1)}, W(0, s) = 0, c(s) = 0, w(x, t) = x(t - 1 + e^{-t})$

7. $w = f(x)g(t), xf'g + f\dot{g} = xt$, take $f(x) = x$ to get $g = ce^{-t} + t - 1$ and $c = 1$ from $w(x, 0) = x(c - 1) = 0$.

11. Set $x^2/(4c^2\tau) = z^2$. Use z as a new variable of integration. Use $\operatorname{erf}(\infty) = 1$.

Chapter 12 Review Questions and Problems, page 603

17. $u = c_1(x)e^{-3y} + c_2(x)e^{2y} - 3$

19. Hyperbolic, $f_1(x) + f_2(y + x)$

21. Hyperbolic, $f_1(y + 2x) + f_2(y - 2x)$

23. $\frac{3}{4}\cos 2t \sin x - \frac{1}{4}\cos 6t \sin 3x$

25. $\sin 0.01\pi x\, e^{-0.001143t}$

27. $\frac{3}{4}\sin 0.01\pi x\, e^{-0.001143t} - \frac{1}{4}\sin 0.03\pi x\, e^{-0.01029t}$

29. $100\cos 2x\, e^{-4t}$

39. $u = (u_1 - u_0)(\ln r)/\ln(r_1/r_0) + (u_0\ln r_1 - u_1\ln r_0)/\ln(r_1/r_0)$

Problem Set 13.1, page 612

1. $1/i = i/i^2 = -i, \quad 1/i^3 = i/i^4 = i$

3. $4.8 - 1.4i$

5. $x - iy = -(x + iy), \quad x = 0$

9. $-117, \quad 4$

11. $-8 - 6i$

13. $-120 - 40i$

15. $3 - i$

17. $-4x^2y^2$

19. $(x^2 - y^2)/(x^2 + y^2), \quad 2xy/(x^2 + y^2)$

Problem Set 13.2, page 618

1. $\sqrt{2}\,(\cos\frac{1}{4}\pi + i\sin\frac{1}{4}\pi)$

3. $2(\cos\frac{1}{2}\pi + i\sin\frac{1}{2}\pi), \quad 2(\cos\frac{1}{2}\pi - i\sin\frac{1}{2}\pi)$

5. $\frac{1}{2}(\cos \pi + i \sin \pi)$

7. $\sqrt{1 + \frac{1}{4}\pi^2}\,(\cos \arctan \frac{1}{2}\pi + i \sin \arctan \frac{1}{2}\pi)$

9. $3\pi/4$

11. $\pm\arctan(\frac{4}{3}) = \pm 0.9273$

13. -1024. *Answer:* π

15. $-3i$

17. $2 + 2i$

21. $\sqrt[6]{2}\,(\cos \frac{1}{12}k\pi + i \sin \frac{1}{12}\pi), \quad k = 1, 9, 17$

23. $6, \quad -3 \pm 3\sqrt{3}i$

25. $\cos(\frac{1}{8}\pi + \frac{1}{2}k\pi) + i \sin(\frac{1}{8}\pi + \frac{1}{2}k\pi), \quad k = 0, 1, 2, 3$

27. $\cos \frac{1}{5}\pi \pm i \sin \frac{1}{5}\pi, \quad \cos \frac{3}{5}\pi \pm i \sin \frac{3}{5}\pi, \quad -1$

29. $i, \quad -1 - i$

31. $\pm(1 - i), \quad \pm(2 + 2i)$

33. $|z_1 + z_2|^2 = (z_1 + z_2)\overline{(z_1 + z_2)} = (z_1 + z_2)(\bar{z}_1 + \bar{z}_2)$. Multiply out and use $\text{Re}\, z_1\bar{z}_2 \leqq |z_1\bar{z}_2|$ (Prob. 34).
$z_1\bar{z}_1 + z_1\bar{z}_2 + z_2\bar{z}_1 + z_2\bar{z}_2 = |z_1|^2 + 2\,\text{Re}\, z_1\bar{z}_2 + |z_2|^2 \leqq |z_1|^2 + 2|z_1||z_2| + |z_2|^2 = (|z_1| + |z_2|)^2$. Hence $|z_1 + z_1|^2 \leqq (|z_1| + |z_2|)^2$. Taking square roots gives (6).

35. $[(x_1 + x_2)^2 + (y_1 + y_2)^2] + [(x_1 - x_2)^2 + (y_1 - y_2)^2] = 2(x_1^2 + y_1^2 + x_2^2 + y_2^2)$

Problem Set 13.3, page 624

1. Closed disk, center $-1 + 5i$, radius $\frac{3}{2}$

3. Annulus (circular ring), center $4 - 2i$, radii π and 3π

5. Domain between the bisecting straight lines of the first quadrant and the fourth quadrant.

7. Half-plane extending from the vertical straight line $x = -1$ to the right.

11. $u(x, y) = (1 - x)/((1 - x)^2 + y^2), \quad u(1, -1) = 0,$
$v(x, y) = y((1 - x)^2 + y^2), \quad v(1, -1) = -1$

15. Yes, since $\text{Im}(|z|^2/z) = \text{Im}(|z|^2\bar{z}/(z\bar{z})) = \text{Im}\,\bar{z} = -r \sin\theta \to 0$.

17. Yes, because $\text{Re}\, z = r\cos\theta \to 0$ and $1 - |z| \to 1$ as $r \to 0$.

19. $f'(z) = 8(z - 4i)^7$. Now $z - 4i = 3$, hence $f'(3 + 4i) = 8 \cdot 3^7 = 17{,}496$.

21. $n(1 - z)^{-n-1}i, \quad ni$

23. $3iz^2/(z + i)^4, \quad -3i/16$

Problem Set 13.4, page 629

1. $r_x = x/r = \cos\theta, \quad r_y = \sin\theta, \quad \theta_x = -(\sin\theta)/r, \quad \theta_y = (\cos\theta)/r$
(a) $0 = u_x - v_y = u_r \cos\theta + u_\theta(-\sin\theta)/r - v_r \sin\theta - v_\theta(\cos\theta)/r$
(b) $0 = u_y + v_x = u_r \sin\theta + u_\theta(\cos\theta)/r + v_r \cos\theta + v_\theta(-\sin\theta)/r$
Multiply (a) by $\cos\theta$, (b) by $\sin\theta$, and add. Etc.

3. Yes

5. No, $f(z) = \overline{(z^2)}$

7. Yes, when $z \neq 0$. Use (7).

9. Yes, when $z \neq 0, \quad -2\pi i, \quad 2\pi i$

11. Yes

13. $f(z) = -\frac{1}{2}i(z^2 + c)$, c real

15. $f(z) = 1/z + c$ (c real)

17. $f(z) = z^2 + z + c$ (c real)

19. No

21. $a = \pi, \quad v = e^{\pi x} \sin \pi y$

23. $a = 0, \quad v = \frac{1}{2}b(y^2 - x^2) + c$

27. $f = u + iv$ implies $if = -v + iu$.

29. Use (4), (5), and (1).

Problem Set 13.5, page 632

3. $e^{2\pi i}e^{-2\pi} = e^{-2\pi} = 0.001867$

5. $e^2(-1) = -7.389$

7. $e^{\sqrt{2}}i = 4.113i$

9. $5e^{i \arctan(3/4)} = 5e^{0.644i}$

11. $6.3e^{\pi i}$

13. $\sqrt{2}e^{\pi i/4}$

15. $\exp(x^2 - y^2)\cos 2xy, \quad \exp(x^2 - y^2)\sin 2xy$
17. $\text{Re}(\exp(z^3)) = \exp(x^3 - 3xy^2)\cos(3x^2y - y^3)$
19. $z = 2n\pi i, \quad n = 0, 1, \cdots$

Problem Set 13.6, page 636

1. Use (11), then (5) for e^{iy}, and simplify.
7. $\cosh 1 = 1.543$, $i \sinh 1 = 1.175i$
9. Both $-0.642 - 1.069i$. Why?
11. $i \sinh \pi = 11.55i$, both
15. Insert the definitions on the left, multiply out, and simplify.
17. $z = \pm(2n + 1)i/2$
19. $z = \pm n\pi i$

Problem Set 13.7, page 640

5. $\ln 11 + \pi i$
7. $\frac{1}{2}\ln 32 - \pi i/4 = 1.733 - 0.785i$
9. $i \arctan(0.8/0.6) = 0.927i$
11. $\ln e + \pi i/2 = 1 + \pi i/2$
13. $\pm 2n\pi i, \quad n = 0, 1, \cdots$
15. $\ln |e^i| + i \arctan \dfrac{\sin 1}{\cos 1} \pm 2n\pi i = 0 + i + 2n\pi i, \quad n = 0, 1, \cdots$
17. $\ln(i^2) = \ln(-1) = (1 \pm 2n)\pi i, \quad 2 \ln i = (1 \pm 4n)\pi i, n = 0, 1, \cdots$
19. $e^{4-3i} = e^4(\cos 3 - i \sin 3) = -54.05 - 7.70i$
21. $e^{0.6}e^{0.4i} = e^{0.6}(\cos 0.4 + i \sin 0.4) = 1.678 + 0.710i$
23. $e^{(1-i)\,\text{Ln}\,(1+i)} = e^{\ln\sqrt{2} + \pi i/4 - i \ln\sqrt{2} + \pi/4} = 2.8079 + 1.3179i$
25. $e^{(3-i)(\ln 3 + \pi i)} = 27e^{\pi}(\cos(3\pi - \ln 3) + i \sin(3\pi - \ln 3)) = -284.2 + 556.4i$
27. $e^{(2-i)\,\text{Ln}(-1)} = e^{(2-i)\pi i} = e^{\pi} = 23.14$

Chapter 13 Review Questions and Problems, page 641

1. $2 - 3i$
3. $27.46e^{0.9929i}, \quad 7.616e^{1.976i}$
11. $-5 + 12i$
13. $0.16 - 0.12i$
15. i
17. $4\sqrt{2}e^{-3\pi i/4}$
19. $15e^{-\pi i/2}$
21. $\pm 3, \quad \pm 3i$
23. $(\pm 1 \pm i)/\sqrt{2}$
25. $f(z) = -iz^2/2$
27. $f(z) = e^{-2z}$
29. $f(z) = e^{-z^2/2}$
31. $\cos 3 \cosh 1 + i \sin 3 \sinh 1 = -1.528 + 0.166i$
33. $i \tanh 1 = 0.7616i$
35. $\cosh \pi \cos \pi + i \sinh \pi \sin \pi = -11.592$

Problem Set 14.1, page 651

1. Straight segment from (2, 1) to (5, 2.5).
3. Parabola $y = x^2$ from (1, 2) to (2, 8).
5. Circle through (0, 0), center $(3, -1)$, radius $\sqrt{10}$, oriented clockwise.
7. Semicircle, center 2, radius 4.
9. Cubic parabola $y = x^3 \quad (-2 \leqq x \leqq 2)$
11. $z(t) = t + (2 + t)i \quad (-1 \leqq t \leqq 1)$
13. $z(t) = 2 - i + 2e^{it} \quad (0 \leqq t \leqq \pi)$

15. $z(t) = 2\cosh t + i\sinh t\ (-\infty < t < \infty)$
17. Circle $z(t) = -a - ib + re^{-it} \quad (0 \leqq t \leqq 2\pi)$
19. $z(t) = t + (1 - \frac{1}{4}t^2)i \quad (-2 \leqq t \leqq 2)$
21. $z(t) = (1 + i)t \quad (1 \leqq t \leqq 3)$, $\operatorname{Re} z = t$, $z'(t) = 1 + i$. *Answer:* $4 + 4i$
23. $e^{2\pi i} - e^{\pi i} = 1 - (-1) = 2$
25. $\frac{1}{2}\exp z^2\big|_1^i = \frac{1}{2}(e^{-1} - e^1) = -\sinh 1$
27. $\tan \frac{1}{4}\pi i - \tan \frac{1}{4} = i \tanh \frac{1}{4} - 1$
29. $\operatorname{Im} z^2 = 2xy = 0$ on the axes. $z = 1 + (-1 + i)t \quad (0 \leqq t \leqq 1)$,
$(\operatorname{Im} z^2)\,\dot{z} = 2(1 - t)y(-1 + i)$ integrated: $(-1 + i)/3$.
35. $|\operatorname{Re} z| = |x| \leqq 3 = M$ on C, $L = \sqrt{8}$

Problem Set 14.2, page 659

1. Use (12), Sec. 14.1, with $m = 2$.
3. Yes
5. 5
7. **(a)** Yes. **(b)** No, we would have to move the contour across $\pm 2i$.
9. 0, yes
11. πi, no
13. 0, yes
15. $-\pi$, no
17. 0, no
19. 0, yes
21. $2\pi i$
23. $1/z + 1/(z - 1)$, hence $2\pi i + 2\pi i = 4\pi i$.
25. 0 (Why?)
27. 0 (Why?)
29. 0

Problem Set 14.3, page 663

1. $2\pi i z^2/(z - 1)\big|_{z=-1} = -\pi i$
3. 0
5. $2\pi i(\cos 3z)/6\big|_{z=0} = \pi i/3$
7. $2\pi i(i/2)^3/2 = \pi/8$
11. $2\pi i \cdot \left.\dfrac{1}{z + 2i}\right|_{z=2i} = \dfrac{\pi}{2}$
13. $2\pi i(z + 2)\big|_{z=2} = 8\pi i$
15. $2\pi i \cosh(-\pi^2 - \pi i) = -2\pi i \cosh \pi^2 = -60{,}739i$ since $\cosh \pi i = \cos \pi = -1$ and $\sinh \pi i = i \sin \pi = 0$.
17. $2\pi i \left.\dfrac{\operatorname{Ln}(z + 1)}{z + i}\right|_{z=i} = 2\pi i \dfrac{\operatorname{Ln}(1 + i)}{2i} = \pi(\ln\sqrt{2} + i\pi/4) = 1.089 + 2.467i$
19. $2\pi i e^{2i}/(2i) = \pi e^{2i}$

Problem Set 14.4, page 667

1. $(2\pi i/3!)(-\cos 0) = -\pi i/3$
3. $(2\pi i/(n - 1)!)e^0$
5. $\dfrac{2\pi i}{3!}(\cosh 2z)''' = \dfrac{\pi i}{3} \cdot 8 \sinh 1 = 9.845i$
7. $(2\pi i/(2n)!)\,(\cos z)^{(2n)}\big|_{z=0} = (2\pi i/(2n)!)(-1)^n \cos 0 = (-1)^n 2\pi i/(2n)!$
9. $-2\pi i(\tan \pi z)'\Big|_{z=0} = \left.\dfrac{-2\pi i \cdot \pi}{\cos^2 \pi z}\right|_{z=0} = -2\pi^2 i$
11. $\dfrac{2\pi i}{4}((1 + z)\sin z)'\Big|_{z=1/2} = \frac{1}{2}\pi i(\sin z + (1 + z)\cos z)\big|_{z=1/2}$
$= \frac{1}{2}\pi i(\sin \frac{1}{2} + \frac{3}{2}\cos \frac{1}{2})$
$= 2.821i$

13. $2\pi i \cdot \frac{1}{z}\Big|_{z=2} = \pi i$

15. 0. Why?

17. 0 by Cauchy's integral theorem for a doubly connected domain; see (6) in Sec. 14.2.

19. $(2\pi i/2!)4^{-3}(e^{3z})''|_{z=\pi i/4} = -9\pi(1 + i)/(64\sqrt{2})$

Chapter 14 Review Questions and Problems, page 668

21. $\frac{1}{2}\cosh(-\frac{1}{4}\pi^2) - \frac{1}{2} = 2.469$

23. $2\pi i(e^z)^{(4)}|_{z=0} = ie^z/12|_{z=0} = \pi i/12$ by Cauchy's integral formula.

25. $-2\pi i(\tan \pi z)'|_{z=1} = -2\pi^2 i/\cos^2 \pi z|_{z=1} = -2\pi^2 i$

27. 0 since $z^2 + \bar{z} - 2 = 2(x^2 - y^2)$ and $y = x$

29. $-4\pi i$

Problem Set 15.1, page 679

1. $z_n = (2i/2)^n$; bounded, divergent, ± 1, $\pm i$

3. $z_n = -\frac{1}{2}\pi i/(1 + 2/(ni))$ by algebra; convergent to $-\pi i/2$

5. Bounded, divergent, $\pm 1 + 10i$

7. Unbounded, hence divergent

9. Convergent to 0, hence bounded

17. Divergent; use $1/\ln n > 1/n$.

19. Convergent; use $\Sigma 1/n^2$.

21. Convergent

23. Convergent

25. Divergent

29. By absolute convergence and Cauchy's convergence principle, for given $\epsilon > 0$ we have for every $n > N(\epsilon)$ and $p = 1, 2, \cdots$

$$|z_{n+1}| + \cdots + |z_{n+p}| < \epsilon,$$

hence $|z_{n+1} + \cdots + z_{n+p}| < \epsilon$ by (6*), Sec. 13.2, hence convergence by Cauchy's principle.

Problem Set 15.2, page 684

1. No! Nonnegative integer powers of z (or $z - z_0$) only!

3. At the center, in a disk, in the whole plane

5. $\Sigma a_n z^{2n} = \Sigma a_n (z^2)^n$, $|z^2| < R = \lim |a_n/a_{n+1}|$; hence $|z| < \sqrt{R}$.

7. $\pi/2, \infty$

9. $i, \sqrt{3}$

11. $0, \sqrt{\frac{26}{5}}$

13. $-i, \frac{1}{2}$

15. $2i$, 1

17. $1/\sqrt{2}$

Problem Set 15.3, page 689

3. $f = \sqrt[n]{n}$. Apply l'Hôpital's rule to $\ln f = (\ln n)/n$.

5. 2

7. $\sqrt{3}$

9. $1/\sqrt{2}$

11. $\sqrt{\frac{7}{3}}$

13. 1

15. $\frac{3}{4}$

Problem Set 15.4, page 697

3. $2z^2 - \frac{(2z^2)^3}{3!} + \cdots = 2z^2 - \frac{4}{3}z^6 + \frac{4}{15}z^{10} - + \cdots, \quad R = \infty$

5. $\frac{1}{2} - \frac{1}{4}z^4 + \frac{1}{8}z^8 - \frac{1}{16}z^{12} + \frac{1}{32}z^{16} - + \cdots, \quad R = \sqrt[4]{2}$

7. $\dfrac{1}{2} + \dfrac{1}{2}\cos z = 1 - \dfrac{1}{2 \cdot 2!}z^2 + \dfrac{1}{2 \cdot 4!}z^4 - \dfrac{1}{2 \cdot 6!}z^6 + - \cdots, \quad R = \infty$

9. $\displaystyle\int_0^z \left(1 - \frac{1}{2}t^2 + \frac{1}{8}t^4 - + \cdots\right) dt = z - \frac{1}{6}z^3 + \frac{1}{40}z^5 - + \cdots, \quad R = \infty$

11. $z^3/(1!3) - z^7/(3!7) + z^{11}/(5!11) - + \cdots, \quad R = \infty$

13. $(2/\sqrt{\pi})(z - z^3/3 + z^5/(2!5) - z^7/(3!7) + \cdots), \quad R = \infty$

17. Team Project. (a) $(\text{Ln}\,(1 + z))' = 1 - z + z^2 - + \cdots = 1/(1 + z)$.

(c) Use that the terms of $(\sin iy)/(iy)$ are all positive, so that the sum cannot be zero.

19. $\frac{1}{2} + \frac{1}{2}i + \frac{1}{2}i(z - i) + (-\frac{1}{4} + \frac{1}{4}i)(z - i)^2 - \frac{1}{4}(z - i)^3 + \cdots, \quad R = \sqrt{2}$

21. $1 - \dfrac{1}{2!}\left(z - \dfrac{1}{2}\pi\right)^2 + \dfrac{1}{4!}\left(z - \dfrac{1}{2}\pi\right)^4 - \dfrac{1}{6!}\left(z - \dfrac{1}{2}\pi\right)^6 + - \cdots, \quad R = \infty$

23. $-\frac{1}{4} - \frac{2}{8}i(z - i) + \frac{3}{16}(z - i)^2 + \frac{4}{32}i(z - i)^3 - \frac{5}{64}(z - i)^4 + \cdots, \quad R = 2$

25. $2\left(z - \dfrac{1}{2}i\right) + \dfrac{2^3}{3!}\left(z - \dfrac{1}{2}i\right)^3 + \dfrac{2^5}{5!}\left(z - \dfrac{1}{2}i\right)^5 + \cdots, \quad R = \infty$

Problem Set 15.5, page 704

3. $|z + i| \leqq \sqrt{3} - \delta, \quad \delta > 0$

5. $|z + \frac{1}{2}i| \leqq \frac{1}{4} - \delta, \quad \delta > 0$

7. Nowhere

9. $|z - 2i| \leqq 2 - \delta, \quad \delta > 0$

11. $|z^n| \leqq 1$ and $\Sigma 1/n^2$ converges. Use Theorem 5.

13. $|\sin^n |z|| \leqq 1$ for all z, and $\Sigma 1/n^2$ converges. Use Theorem 5.

15. $R = 4$ by Theorem 2 in Sec. 15.2; use Theorem 1.

17. $R = 1/\sqrt{\pi} > 0.56$; use Theorem 1.

Chapter 15 Review Questions and Problems, page 706

11. 1

13. 3

15. $\frac{1}{2}$

17. $\infty, \quad e^{2z}$

19. $\infty, \quad \cosh \sqrt{z}$

21. $\displaystyle\sum_{n=0}^{\infty} \frac{z^{4n}}{(2n + 1)!}, \quad R = \infty$

23. $\displaystyle\frac{1}{2} + \frac{1}{2}\cos 2z = 1 + \frac{1}{2}\sum_{n=1}^{\infty} \frac{(-1)^n}{(2n)!}(2z)^{2n}, \quad R = \infty$

25. $\displaystyle\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n!} z^{2n-2}, \quad R = \infty$

27. $\cos\,[(z - \frac{1}{2}\pi) + \frac{1}{2}\pi] = -(z - \frac{1}{2}\pi) + \frac{1}{6}(z - \frac{1}{2}\pi)^3 - + \cdots = -\sin\,(z - \frac{1}{2}\pi)$

29. $\ln 3 + \dfrac{1}{3}(z - 3) - \dfrac{1}{2 \cdot 9}(z - 3)^2 + \dfrac{1}{3 \cdot 27}(z - 3)^3 - + \cdots, \quad R = 3$

Problem Set 16.1, page 714

1. $z^{-4} - \frac{1}{2}z^{-2} + \frac{1}{24} - \frac{1}{720}z^2 + - \cdots, \quad 0 < |z| < \infty$

3. $z^{-3} + z^{-1} + \frac{1}{2}z + \frac{1}{6}z^3 + \frac{1}{24}z^5 + \cdots, \quad 0 < |z| < \infty$

5. $z^{-2} + z^{-1} + 1 + z + z^2 + \cdots, \quad 0 < |z| < 1$

7. $z^3 + \frac{1}{2}z + \frac{1}{24}z^{-1} + \frac{1}{720}z^3 + \cdots, \quad 0 < |z| < \infty$

9. $\exp[1 + (z - 1)](z - 1)^{-2} = e \cdot [(z - 1)^{-2} + (z - 1)^{-1} + \frac{1}{2} + \frac{1}{6}(z - 1) + \cdots],$
$0 < |z - 1| < \infty$

11. $$\frac{[\pi i + (z - \pi i)]^2}{(z - \pi i)^4} = \frac{(\pi i)^2}{(z - \pi i)^4} + \frac{2\pi i}{(z - \pi i)^3} + \frac{1}{(z - \pi i)^2}$$

13. $$i^{-3}\left(1 + \frac{z - i}{i}\right)^{-3}(z - i)^{-2} = \sum_{n=0}^{\infty}\binom{-3}{n} i^{-3-n}(z - i)^{n-2} = i(z - i)^{-2}$$
$-3(z - i)^{-1} - 6i + 10(z - i) + \cdots, \quad 0 < |z - i| < 1$

15. $(-\cos(z - \pi))(z - \pi)^{-2} = -(z - \pi)^{-2} + \frac{1}{2} - \frac{1}{24}(z - \pi)^2 + - \cdots,$
$0 < |z - \pi| < \infty$

19. $$\sum_{n=0}^{\infty} z^{2n}, \quad |z| < 1, \qquad -\sum_{n=0}^{\infty} \frac{1}{z^{2n+2}}, \quad |z| > 1$$

21. $-(z + \frac{1}{2}\pi)^{-1}\cos(z + \frac{1}{2}\pi) = -(z + \frac{1}{2}\pi)^{-1} + \frac{1}{2}(z + \frac{1}{2}\pi) - \frac{1}{24}(z + \frac{1}{2}\pi)^3 + - \cdots,$
$|z + \frac{1}{2}\pi| > 0$

23. $z^8 + z^{12} + z^{16} + \cdots, \quad |z| < 1, \quad -z^4 - 1 - z^{-4} - z^{-8} - \cdots, \quad |z| > 1$

25. $$\frac{i}{(z - i)^2} + \frac{1}{z - i} + i + (z - i)$$

Section 16.2, page 719

1. $0 \pm 2\pi, \pm 4\pi, \cdots$, fourth order

3. $-81i$, fourth order

5. $\pm 1, \pm 2, \cdots$, second order

7. $\pm(2 + 2i), \pm i$, simple

9. $\frac{1}{2}\sin 4z$, $z = 0, \pm\pi/4, \pm\pi/2, \cdots$, simple

11. $f(z) = (z - z_0)^n g(z)$, $g(z_0) \neq 0$, hence $f^2(z) = (z - z_0)^{2n} g^2(z)$.

13. Second-order poles at i and $-2i$

15. Simple pole at ∞, essential singularity at $1 + i$

17. Fourth-order poles at $\pm n\pi i$, $n = 0, 1, \cdots$, essential singularity at ∞

19. $e^z(1 - e^z) = 0$, $e^z = 1$, $z = \pm 2n\pi i$ simple zeros. *Answer:* simple poles at $\pm 2n\pi i$, essential singularity at ∞

21. $1, \infty$ essential singularities, $\pm 2n\pi i$, $n = 0, 1, \cdots$, simple poles

Section 16.3, page 725

3. $\frac{4}{15}$ at 0

5. $\pm 4i$ at $\mp i$

7. $1/\pi$ at $0, \pm 1, \cdots$

9. -1 at $\pm 2n\pi i$

11. $(e^z)''/2!|_{z=\pi i} = -\frac{1}{2}$ at $z = \pi i$

15. Simple pole at $\frac{1}{4}$ inside C, residue $-1/(2\pi)$. *Answer:* $-i$

17. Simple poles at $\pi/2$, residue $e^{\pi/2}/(-\sin \pi/2)$, and at $-\pi/2$, residue $e^{-\pi/2}/\sin \pi/2 = e^{-\pi/2}$. *Answer:* $-4\pi i \sinh \pi/2$

19. $2\pi i(\sinh \frac{1}{2}i)/2 = -\pi \sin \frac{1}{2}$

21. $z^{-5}\cos \pi z = \cdots + \pi^4/(4!z) - + \cdots$. *Answer:* $2\pi^5 i/24$

23. Residues $\frac{1}{2}$ at $z = \frac{1}{2}$, 2 at $z = \frac{1}{3}$. *Answer:* $5\pi i$

25. Simple poles inside C at $2i, -2i, 3i, -3i$, residues $(2i \cosh 2i)/(4z^3 + 26z)|_{z=2i} = \frac{1}{10}, \frac{1}{10}, \frac{1}{10}, \frac{1}{10}$, respectively. *Answer:* $2\pi i \cdot \frac{4}{10}$

Problem Set 16.4, page 733

1. $2\pi/\sqrt{k^2 - 1}$

3. $\pi/\sqrt{2}$

5. $5\pi/12$

7. $2a\pi/\sqrt{a^2 - 1}$

9. 0. Why? (Make a sketch.)

11. $\pi/2$

13. 0. Why?

15. $\pi/3$

17. 0. Why?

19. Simple poles at $\pm 1, i$ (and $-i$); $2\pi i \cdot \frac{1}{4}i + \pi i(-\frac{1}{4} + \frac{1}{4}) = -\frac{1}{2}\pi$

21. Simple poles at 1 and $\pm 2\pi i$, residues i and $-i$. *Answer:* $\dfrac{\pi}{5}(\cos 1 - e^{-2})$

23. $-\pi/2$

25. 0

27. Let $q(z) = (z - a_1)(z - a_2) \cdots (z - a_k)$. Use (4) in Sec. 16.3 to form the sum of the residues $1/q'(a_1) + \cdots + 1/q'(a_k)$ and show that this sum is 0; here $k > 1$.

Chapter 16 Review Questions and Problems, page 733

11. $6\pi i$

13. $2\pi i(-10 - 10)$

15. $2\pi i(25z^2)'|_{z=5} = 500\pi i$

17. 0 (n even), $(-1)^{(n-1)/2} 2\pi i/(n - 1)!$ (n odd)

19. $\pi/6$

21. $\pi/60$

23. 0. Why?

25. $\operatorname{Res}_{z=i} e^{iz}/(z^2 + 1) = 1/(2ie)$. *Answer:* π/e.

Problem Set 17.1, page 741

5. Only in size

7. $x = c, w = -y + ic;\quad y = k, w = -k + ix$

9. Parallel displacement; each point is moved 2 to the right and 1 up.

11. $|w| \leqq \frac{1}{4}, \quad -\pi/4 < \operatorname{Arg} w < \pi/4$

13. $-5 \leqq \operatorname{Re} z \leqq -2$

15. $u \geqq 1$

17. Annulus $\frac{1}{2} \leqq |w| \leqq 4$

19. $0 < u < \ln 4, \quad \pi/4 < v \leqq 3\pi/4$

21. $z^3 + az^2 + bz + c, \quad z = -\frac{1}{3}(a \pm \sqrt{a^2 - 3b})$

23. $z = (-1 \pm \sqrt{3})/2$

25. $\sinh z = 0$ at $z = 0, \quad \pm\pi i, \pm 2\pi i, \cdots$

29. $M = |z| = 1$ on the unit circle, $J = |z|^2$

31. $|w'| = 1/|z|^2 = 1$ on the unit circle, $J = 1/|z|^4$

33. $M = e^x = 1$ for $x = 0$, the y-axis, $J = e^{2x}$

35. $M = 1/|z| = 1$ on the unit circle, $J = 1/|z|^2$

Problem Set 17.2, page 745

7. $z = \dfrac{w + i}{2w}$

9. $z = \dfrac{4w + i}{-3iw + 1}$

11. $z = 0, \quad 1/(a + ib)$

13. $z = 0, \quad \pm\frac{1}{2}, \pm = \pm i/2$

15. $z = i, 2i$ **17.** $w = \dfrac{az}{cz + a}$ **19.** $w = \dfrac{az + b}{-bz + a}$

Problem Set 17.3, page 750

3. Apply the inverse g of f on both sides of $z_1 = f(z_1)$ to get $g(z_1) = g(f(z_1)) = z_1$.
9. $w = iz$, a rotation. Sketch to see.
11. $w = (z + i)/(z - i)$
13. $w = 1/z$, almost by inspection
15. $w = 1/z - 1$
17. $w = (2z - i)/(-iz - 2)$
19. $w = (z^4 - i)(-iz^4 + 1)$

Problem Set 17.4, page 754

1. Circle $|w| = e^c$
3. Annulus $1/\sqrt{e} \leqq |w| \leqq \sqrt{e}$
5. w-plane without $w = 0$
7. $1 < |w| < e, v > 0$
9. $\pm(2n + 1)\pi/2, \quad n = 0, 1, \cdots$
11. $u^2/\cosh^2 2 + v^2/\sinh^2 2 < 1, \quad u > 0, v > 0$
13. Elliptic annulus bounded by $u^2/\cosh^2 1 + v^2/\sinh^2 1 = 1$ and $u^2/\cosh^2 3 + v^2/\sinh^2 3 = 1$
15. $\cosh z = \cos iz = \sin (iz + \frac{1}{2}\pi)$
17. $0 < \text{Im}\, t < \pi$ is the image of R under $t = z^2/2$. *Answer:* $e^t = e^{z^2/2}$.
19. Hyperbolas $u^2/\cos^2 c - v^2/\sin^2 c = \cosh^2 c - \sinh^2 c = 1$ when $c \neq 0, \pi$, and $u = \pm\cosh y$ (thus $|u| \geqq 1$), $v = 0$ when $c = 0, \pi$.
21. Interior of $u^2/\cosh^2 2 + v^2/\sinh^2 2 = 1$ in the fourth quadrant, or map $\pi/2 < x < \pi, 0 < y < 2$ by $w = \sin z$ (why?).
23. $v < 0$
25. The images of the five points in the figure can be obtained directly from the function w.

Problem Set 17.5, page 756

1. w moves once around the circle $|w| = \frac{1}{2}$.
3. Four sheets, branch point at $z = -1$
5. $-i/4$, three sheets
7. z_0, n sheets
9. $\sqrt{z(z - i)(z + i)}$, $0, \pm i$, two sheets

Chapter 17 Review Questions and Problems, page 756

11. $1 < |w| < 4, |\arg w| < \pi/4$
13. Horizontal strip $-8 < v < 8$
15. $u = 1 - \frac{1}{4}v^2$, same (why?)
17. $|w| > 1$
19. $\frac{1}{3} < |w| < \frac{1}{2}, \quad v < 0$
21. $w = 1 + iv, \quad v < 0$
23. $w = \dfrac{10z + 5i}{z + 2i}$
25. Rotation $w = iz$
27. $w = 1/z$
29. $z = 0$
31. $z = 2 \pm \sqrt{6}$
33. $z = 0, \pm i, \pm 3i$
35. $w = e^{4z}$
37. $w = iz^2 + 1$
39. $w = z^2/(2c)$

Problem Set 18.1, page 762

1. 2.5 mm = 0.25 cm; $\Phi = \text{Re}\, 110(1 + (\text{Ln}\, z)/\ln 4)$

3. $\Phi = \text{Re}\left(30 - \dfrac{20}{\ln 10}\text{Ln}\, z\right)$

5. $\Phi(x) = \text{Re}\,(375 + 25z)$

7. $\Phi(r) = \text{Re}\,(32 - z)$

13. Use Fig. 391 in Sec. 17.4 with the z- and w-planes interchanged and $\cos z = \sin(z + \frac{1}{2}\pi)$.

15. $\Phi = 220(x^3 - 3xy^2) = \text{Re}\,(220z^3)$

Problem Set 18.2, page 766

3. $w = iz^2$ maps R onto the strip $-2 \leqq u \leqq 0$; and $\Phi^* = U_2 + (U_1 - U_2)(1 + \frac{1}{2}u) = U_2 + (U_1 - U_2)(1 - xy)$.

5. **(a)** $\dfrac{(x-2)(2x-1) + 2y^2}{(x-2)^2 + y^2} = c$, **(b)** $x^2 - y^2 = c$, $xy = c$, $e^x \cos y = c$

7. See Fig. 392 in Sec. 17.4. $\Phi = \text{Re}\,(\sin^2 z)$, $\sin^2 x\ (y = 0)$, $\sin^2 x \cosh^2 1 - \cos^2 x \sinh^2 1\ (y = 1)$, $-\sinh^2 y\ (x = 0, \pi)$.

9. $\Phi(x, y) = \cos^2 x \cosh^2 y - \sin^2 x \sinh^2 y$; $\cosh^2 y\ (x = 0)$, $-\sinh y\ (x = \frac{\pi}{2})$, $\cos^2 x\ (y = 0)$, $\cos^2 x \cosh^2 1 - \sin^2 x \sinh^2 1\ (y = 1)$

13. Corresponding rays in the w-plane make equal angles, and the mapping is conformal.

15. Apply $w = z^2$.

17. $z = (2Z - i)/(-iZ - 2)$ by (3) in Sec. 17.3.

19. $\Phi = \dfrac{5}{\pi}\text{Arg}\,(z - 2)$, $F = -\dfrac{5i}{\pi}\text{Ln}\,(z - 2)$

Problem Set 18.3, page 769

1. $(80/d)y + 20$. Rotate through $\pi/2$.

5. $\dfrac{80}{\pi}\arctan\dfrac{y}{x} = \text{Re}\left(-\dfrac{80i}{\pi}\text{Ln}\, z\right)$

7. $T_1 + \dfrac{2}{\pi}(T_2 - T_1)\arctan\dfrac{y}{x} = \text{Re}\left(T_1 - \dfrac{2i}{\pi}(T_2 - T_1)\,\text{Ln}\, z\right)$

9. $\dfrac{T_1}{\pi}\left(\arctan\dfrac{y}{x-b} - \arctan\dfrac{y}{x-a}\right) = \text{Re}\left(\dfrac{iT_1}{\pi}\text{Ln}\,\dfrac{z-a}{z-b}\right)$

11. $\dfrac{100}{\pi}(\text{Arg}\,(z-1) - \text{Arg}\,(z+1)) = \text{Re}\left(\dfrac{100i}{\pi}\text{Ln}\,\dfrac{z+1}{z-1}\right)$

13. $\dfrac{100}{\pi}[\text{Arg}\,(z^2 - 1) - \text{Arg}\,(z^2 + 1)]$ from $w = z^2$ and Prob. 11.

15. $-20 + (320/\pi)\,\text{Arg}\, z = \text{Re}\left(-20 - \dfrac{320i}{\pi}\text{Ln}\, z\right)$

17. $\text{Re}\, F(z) = 100 + (200/\pi)\,\text{Re}\,(\arcsin z)$

Problem Set 18.4, page 776

1. $V(z)$ continuously differentiable.

3. $|F'(iy)| = 1 + 1/y^2$, $|y| \geqq 1$, is maximum at $y = \pm 1$, namely, 2.

5. Calculate or note that $\nabla^2 = \text{div grad}$ and curl grad is the zero vector; see Sec. 9.8 and Problem Set 9.7.
7. Horizontal parallel flow to the right.
9. $F(z) = z^4$
11. Uniform parallel flow upward, $V = \overline{F'} = iK$, $V_1 = 0$, $V_2 = K$
13. $F(z) = z^3$
15. $F(z) = z/r_0 + r_0/z$
17. Use that $w = \arccos z$ gives $z = \cos w$ and interchanging the roles of the z- and w-planes.
19. $y/(x^2 + y^2) = c$ or $x^2 + (y - k)^2 = k^2$

Problem Set 18.5, page 781

5. $\Phi = \frac{3}{2} r^3 \sin 3\theta$
7. $\Phi = \frac{1}{2} a + \frac{1}{2} ar^8 \cos 8\theta$
9. $\Phi = 3 - 4r^2 \cos 2\theta + r^4 \cos 4\theta$
11. $\Phi = \dfrac{2}{\pi}\left(r \sin \theta - \dfrac{1}{2} r^2 \sin 2\theta + \dfrac{1}{3} r^3 \sin 3\theta - + \cdots\right)$
13. $\Phi = \dfrac{2}{\pi} r \sin \theta + \dfrac{1}{2} r^2 \sin 2\theta - \dfrac{2}{9\pi} r^3 \sin 3\theta - \dfrac{1}{4} r^4 \sin 4\theta + + - \cdots$
15. $\Phi = \dfrac{1}{2} + \dfrac{2}{\pi}\left(r \cos \theta - \dfrac{1}{3} r^3 \cos 3\theta + \dfrac{1}{5} r^5 \cos 5\theta - + \cdots\right)$
17. $\Phi = \dfrac{1}{3} - \dfrac{4}{\pi^2}\left(r \cos \theta - \dfrac{1}{4} r^2 \cos 2\theta + \dfrac{1}{9} r^3 \cos 3\theta - + \cdots\right)$

Problem Set 18.6, page 784

1. Use (2). $F(z_0 + e^{i\alpha}) = (\frac{7}{2} + e^{i\alpha})^3$, etc. $F(\frac{5}{2}) = \frac{343}{8}$
3. Use (2). $F(z_0 + e^{i\alpha}) = (2 + 3e^{i\alpha})^2$, etc. $F(4) = 100$
5. No, because $|z|$ is not analytic.
7.
$$\Phi(2, -2) = -3 = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} (1 + r \cos \alpha)(-3 + r \sin \alpha) r \, dr \, d\alpha$$
$$= \frac{1}{\pi}\int_0^1 \int_0^{2\pi} (-3r + \cdots) \, dr \, d\alpha = \frac{1}{\pi}\left(-\frac{3}{2}\right) \cdot 2\pi$$
9.
$$\Phi(1, 1) = 3 = \frac{1}{\pi}\int_0^1 \int_0^{2\pi} (3 + r \cos \alpha + r \sin \alpha + r^2 \cos \alpha \sin \alpha) r \, dr \, d\alpha$$
$$= \frac{1}{\pi} \cdot \frac{3}{2} \cdot 2\pi$$
13. $|F(z)| = [\cos^2 x + \sinh^2 y]^{1/2}$, $z = \pm i$, Max $= [1 + \sinh^2 1]^{1/2} = 1.543$
15. $|F(z)|^2 = \sinh^2 2x \cos^2 2y + \cosh^2 2x \sin^2 2y = \sinh^2 2x + 1 \cdot \sin^2 2y$, $z = 1$, Max $= \sinh 2 = 3.627$
17. $|F(z)|^2 = 4(2 - 2 \cos 2\theta)$, $z = \pi/2$, $3\pi/2$, Max $= 4$
19. No. Make up a counterexample.

Chapter 18 Review Questions and Problems, page 785

11. $\Phi = 10(1 - x + y), \quad F = 10 - 10(1 + i)z$

13. $\Phi = \text{Re}\,(220 - 95.54 \text{ Ln } z) = 220 - \dfrac{220}{\ln 10} \ln r = 220 - 95.54 \ln r.$

17. $2(1 - (2/\pi) \text{ Arg } z)$

19. $30(1 - (2/\pi) \text{ Arg } (z - 1))$

21. $\Phi = x + y = \text{const}, \quad V = \overline{F'(z)} = 1 - i, \quad$ parallel flow

23. $T(x, y) = x(2y + 1) = \text{const}$

25. $\overline{F'(z)} = \bar{z} + 1 = x + 1 - iy$

Problem Set 19.1, page 796

1. $0.84175 \cdot 10^2, \quad -0.52868 \cdot 10^3, \quad 0.92414 \cdot 10^{-3}, \quad -0.36201 \cdot 10^6$

3. 6.3698, 6.794, 8.15, impossible

5. Add first, then round.

7. 29.9667, 0.0335; 29.9667, 0.0333704 (6S-exact)

9. 29.97, 0.035; 29.97, 0.03337; 30, 0.0; 30, 0.033

11. $|\epsilon| = |x + y - (\tilde{x} + \tilde{y})| = |(x - \tilde{x}) + (y - \tilde{y})| = |\epsilon_x + \epsilon_y|$
$\leqq |\epsilon_x| + |\epsilon_y| = \beta_x + \beta_y$

13. $\dfrac{a_1}{a_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2 + \epsilon_2} = \dfrac{\tilde{a}_1 + \epsilon_1}{\tilde{a}_2}\left(1 - \dfrac{\epsilon_2}{\tilde{a}_2} + \dfrac{\epsilon_2^2}{\tilde{a}_2^2} - + \cdots\right) \approx \dfrac{\tilde{a}_1}{\tilde{a}_2} + \dfrac{\epsilon_1}{\tilde{a}_2} - \dfrac{\epsilon_2}{\tilde{a}_2} \cdot \dfrac{\tilde{a}_1}{\tilde{a}_2},$

hence $\left|\left(\dfrac{a_1}{a_2} - \dfrac{\tilde{a}_1}{\tilde{a}_2}\right)\Big/\dfrac{a_1}{a_2}\right| \approx \left|\dfrac{\epsilon_1}{a_1} - \dfrac{\epsilon_2}{a_2}\right| \leqq |\epsilon_{r1}| + |\epsilon_{r2}| \leqq \beta_{r1} + \beta_{r2}$

15. **(a)** $1.38629 - 1.38604 = 0.00025$, **(b)** $\ln 1.00025 = 0.000249969$ is 6S-exact.

19. In the present case, (b) is slightly more accurate than (a) (which may produce nonsensical results; cf. Prob. 20).

21. $c_4 \cdot 2^4 + \cdots + c_0 \cdot 2^0 = (1\ 0\ 1\ 1\ 1.)_2$, NOT $(1\ 1\ 1\ 0\ 1.)_2$

23. The algorithm in Prob. 22 repeats 0011 infinitely often.

25. $n = 26$. The beginning is 0.09375 ($n = 1$).

27. $I_{14} = 0.1812$ (0.1705 4S-exact), $\quad I_{13} = 0.1812$ (0.1820), $\quad I_{12} = 0.1951$ (0.1951), $I_{11} = 0.2102$ (0.2103), etc.

29. $-0.126 \cdot 10^{-2}, -0.402 \cdot 10^{-3}; \quad -0.266 \cdot 10^{-6}, -0.847 \cdot 10^{-7}$

Problem Set 19.2, page 807

3. $g = 0.5 \cos x, \quad x = 0.450184$ ($= x_{10}$, exact to 6S)

5. Convergence to 4.7 for all these starting values.

7. $x = x/(e^x \sin x)$; 0.5, 0.63256, $\cdots$ converges to 0.58853 (5S-exact) in 14 steps.

9. $x = x^4 - 0.12; \quad x_0 = 0, x_3 = -0.119794$ (6S-exact)

11. $g = 4/x + x^3/16 - x^5/576; \quad x_0 = 2, x_n = 2.39165$ ($n \geqq 6$), 2.405 4S-exact

13. This follows from the intermediate value theorem of calculus.

15. $x_3 = 0.450184$

17. Convergence to $x = 4.7, 4.7, 0.8, -0.5$, respectively. Reason seen easily from the graph of f.

19. 0.5, 0.375, 0.377968, 0.377964; (b) $1/\sqrt{7}$

21. 1.834243 ($= x_4$), 0.656620 ($= x_4$), −2.49086 ($= x_4$)

23. $x_0 = 4.5$, $x_4 = 4.73004$ (6S-exact)

25. (a) ALGORITHM BISECT (f, a_0, b_0, ϵ, N) Bisection Method

This algorithm computes the solution c of $f(x) = 0$ (f continuous) within the tolerance ϵ, given an initial interval $[a_0, b_0]$ such that $f(a_0)f(b_0) < 0$.

INPUT: Continuous function f, initial interval $[a_0, b_0]$, tolerance ϵ, maximum number of iterations N.

OUTPUT: A solution c (within the tolerance ϵ), or a message of failure.

For $n = 0, 1, \cdots, N - 1$ do:

$c = \frac{1}{2}(a_n + b_n)$

If $f(c) = 0$ then OUTPUT c Stop. [*Procedure completed*]

Else if $f(a_n)f(b_n) < 0$ then set $a_{n+1} = a_n$ and $b_{n+1} = c$.

Else set $a_{n+1} = c$, and $b_{n+1} = b_n$.

If $|a_{n+1} - b_{n+1}| < \epsilon|c|$ then OUTPUT c. Stop. [*Procedure completed*]

End

OUTPUT $[a_N, b_N]$ and a message "Failure". Stop.

[*Unsuccessful completion; N iterations did not give an interval of length not exceeding the tolerance.*]

End BISECT

Note that $[a_N, b_N]$ gives $(a_N + b_N)/2$ as an approximation of the zero and $(b_N - a_N)/2$ as a corresponding error bound.

(b) 0.739085; **(c)** 1.30980, 0.429494

27. $x_2 = 1.5$, $x_3 = 1.76471, \cdots$, $x_7 = 1.83424$ (6S-exact)

29. 0.904557 (6S-exact)

Problem Set 19.3, page 819

1. $L_0(x) = -2x + 19$, $L_1(x) = 2x - 18$, $p_1(9.3) = L_0(9.3) \cdot f_0 + L_1(9.3) \cdot f_1 = 0.1086 \cdot 9.3 + 1.230 = 2.2297$

3. $$p_2(x) = \frac{(x - 1.02)(x - 1.04)}{(-0.02)(-0.04)} \cdot 1.0000 + \frac{(x - 1)(x - 1.04)}{0.02\,(-0.02)} \cdot 0.9888 + \frac{(x - 1)(x - 1.02)}{0.04 \cdot 0.02} \cdot 0.9784 = x^2 - 2.580x + 2.580;$$ 0.9943, 0.9835

5. 0.8033 (error −0.0245), 0.4872 (error −0.0148); quadratic: 0.7839 (−0.0051), 0.4678 (0.0046)

7. $p_2(x) = 1.1640x - 0.3357x^2$; −0.5089 (error 0.1262), 0.4053 (−0.0226), 0.9053 (0.0186), 0.9911 (−0.0672)

9. $p_2(x) = -0.44304x^2 + 1.30896x - 0.023220$, $p_2(0.75) = 0.70929$ (5S-exact 0.71116)

11. $L_0 = -\frac{1}{6}(x - 1)(x - 2)(x - 3)$, $L_1 = \frac{1}{2}x(x - 2)(x - 3)$, $L_2 = -\frac{1}{2}x(x - 1)(x - 3)$, $L_3 = \frac{1}{6}x(x - 1)(x - 2)$; $p_3(x) = 1 + 0.039740x - 0.335187x^2 + 0.060645x^3$; $p_2(0.5) = 0.943654$, $p_3(1.5) = 0.510116$, $p_3(2.5) = -0.047991$

13. $2x^2 - 4x + 2$

15. $p_3(x) = 2.1972 + (x - 9) \cdot 0.1082 + (x - 9)(x - 9.5) \cdot 0.005235$

17. $r = -1.5$, $p_2(0.3) = 0.6039 + (-1.5) \cdot 0.1755 + \frac{1}{2}(-1.5)(-0.5) \cdot (-0.0302) = 0.3293$

Problem Set 19.4, page 826

9. $[-1.39(x-5)^2 + 0.58(x-5)^3]'' = 0.004$ at $x = 5.8$ (due to roundoff; should be 0).

11. $1 - \frac{5}{4}x^2 + \frac{1}{4}x^4$

13. $1 - x^2$, $-2(x-1) - (x-1)^2 + 2(x-1)^3$, $-1 + 2(x-2) + 5(x-2)^2 - 6(x-2)^3$

15. $4 + x^2 - x^3$, $-8(x-2) - 5(x-2)^2 + 5(x-2)^3$, $4 + 32(x-4) + 25(x-4)^2 - 11(x-4)^3$

17. Use the fact that the third derivative of a cubic polynomial is constant, so that g''' is piecewise constant, hence constant throughout under the present assumption. Now integrate three times.

19. Curvature $f''/(1 + f'^2)^{3/2} \approx f''$ if $|f'|$ is small.

Problem Set 19.5, page 839

1. 0.747131, which is larger than 0.746824. Why?

3. 0.5, 0.375, 0.34375, 0.335 (exact)

5. $\epsilon_{0.5} \approx 0.03452$ ($\epsilon_{0.5} = 0.03307$), $\epsilon_{0.25} \approx 0.00829$ ($\epsilon_{0.25} = 0.00820$)

7. 0.693254 (6S-exact 0.693147)

9. 0.073930 (6S-exact 0.073928)

11. 0.785392 (6S-exact 0.785398)

13. $(0.785398126 - 0.785392156)/15 = 0.39792 \cdot 10^{-6}$

15. **(a)** $M_2 = 2$, $|KM_2| = 2/(12n^2) = 10^{-5}/2$, $n = 183$. **(b)** $f^{\text{iv}} = 24/x^5$, $M_4 = 24$, $|CM_4| = 24/(180 \cdot (2m)^4) = 10^{-5}/2$, $2m = 12.8$, hence 14.

17. 0.94614588, 0.94608693 (8S-exact 0.94608307)

19. 0.9460831 (7S-exact)

21. 0.9774586 (7S-exact 0.9774377)

23. Set $x = \frac{1}{2}(t+1)$, 0.2642411177 (10S-exact), $1 - 2/e$

25. $x = \frac{1}{2}(t+1)$, $dx = \frac{1}{2}dt$, 0.746824127 (9S-exact 0.746824133)

27. 0.08, 0.32, 0.176, 0.256 (exact)

29. $5(0.1040 - \frac{1}{2} \cdot 0.1760 + \frac{1}{3} \cdot 0.1344 - \frac{1}{4} \cdot 0.0384) = 0.256$

Chapter 19 Review Questions and Problems, page 841

17. 4.375, 4.50, 6.0, impossible

19. $44.885 \leqq s \leqq 44.995$

21. The same as that of $\tilde{a}$.

23. $x = 20 \pm \sqrt{398} = 20.00 \pm 19.95$, $x_1 = 39.95$, $x_2 = 0.05$, $x_2 = 2/39.95 = 0.05006$ (error less than 1 unit of the last digit)

25. $x = x^4 - 0.1$, -0.1, -0.999, -0.99900399

27. 0.824

29. $-x + x^3$, $2(x-1) + 3(x-1)^2 - (x-1)^3$

31. 0.26, $M_2 = 6$, $M_2^* = 0$, $-0.02 \leqq \epsilon \leqq 0$, 0.01

33. 0.90443, 0.90452 (5S-exact 0.90452)

35. (a) $(0.4^3 - 2 \cdot 0.2^3 + 0)/0.04 = 1.2$, (b) $(0.3^3 - 2 \cdot 0.2^3 + 0.1^3)/0.01 = 1.2$ (exact)

Problem Set 20.1, page 851

1. $x_1 = 7.3, \quad x_2 = -3.2$ **3.** No solution **5.** $x_1 = 2, \quad x_2 = 1$

7. $\begin{bmatrix} -3 & 6 & -9 & -46.725 \\ 0 & 9 & -13 & -51.223 \\ 0 & 0 & -2.88889 & -7.38689 \end{bmatrix}$

$x_1 = 3.908, \quad x_2 = -1.998, \quad x_3 = 2.557$

9. $\begin{bmatrix} 13 & -8 & 0 & 178.54 \\ 0 & 6 & 13 & 137.86 \\ 0 & 0 & -16 & -253.12 \end{bmatrix}$

$x_1 = 6.78, \quad x_2 = -11.3, \quad x_3 = 15.82$

11. $\begin{bmatrix} 3.4 & -6.12 & -2.72 & 0 \\ 0 & 0 & 4.32 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

$x_1 = t_1$ arbitrary, $\quad x_2 = (3.4/6.12)t_1, \quad x_3 = 0$

13. $\begin{bmatrix} 5 & 0 & 6 & -0.329193 \\ 0 & -4 & -3.6 & -2.143144 \\ 0 & 0 & 2.3 & -0.4 \end{bmatrix}$

$x_1 = 0.142856, \quad x_2 = 0.692307, \quad x_3 = -0.173912$

15. $\begin{bmatrix} -1 & -3.1 & 2.5 & 0 & -8.7 \\ 0 & 2.2 & 1.5 & -3.3 & -9.3 \\ 0 & 0 & -1.493182 & -0.825 & 1.03773 \\ 0 & 0 & 0 & 6.13826 & 12.2765 \end{bmatrix}$

$x_1 = 4.2, \quad x_2 = 0, \quad x_3 = -1.8, \quad x_4 = 2.0$

Problem Set 20.2, page 857

1. $\begin{bmatrix} 1 & 0 \\ 3 & 1 \end{bmatrix} \begin{bmatrix} 4 & 5 \\ 0 & -1 \end{bmatrix}, \quad \begin{aligned} x_1 &= -4 \\ x_2 &= 6 \end{aligned}$

3. $\begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & 0 \\ 2 & 5 & 1 \end{bmatrix} \begin{bmatrix} 5 & 4 & 1 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{bmatrix}, \quad \begin{aligned} x_1 &= 0.4 \\ x_2 &= 0.8 \\ x_3 &= 1.6 \end{aligned}$

5. $\begin{bmatrix} 1 & 0 & 0 \\ 6 & 1 & 0 \\ 3 & 9 & 1 \end{bmatrix} \begin{bmatrix} 3 & 9 & 6 \\ 0 & -6 & 3 \\ 0 & 0 & -3 \end{bmatrix}, \quad \begin{aligned} x_1 &= -\tfrac{1}{15} \\ x_2 &= \tfrac{4}{15} \\ x_3 &= \tfrac{2}{5} \end{aligned}$

7. $\begin{bmatrix} 3 & 0 & 0 \\ 2 & 3 & 0 \\ 4 & 1 & 3 \end{bmatrix} \begin{bmatrix} 3 & 2 & 4 \\ 0 & 3 & 1 \\ 0 & 0 & 3 \end{bmatrix}$, $\begin{matrix} x_1 = 0.6 \\ x_2 = 1.2 \\ x_3 = 0.4 \end{matrix}$

9. $\begin{bmatrix} 0.1 & 0 & 0 \\ 0 & 0.4 & 0 \\ 0.3 & 0.2 & 0.1 \end{bmatrix} \begin{bmatrix} 0.1 & 0 & 0.3 \\ 0 & 0.4 & 0.2 \\ 0 & 0 & 0.1 \end{bmatrix}$, $\begin{matrix} x_1 = 2 \\ x_2 = -11 \\ x_3 = 4 \end{matrix}$

11. $\begin{bmatrix} 1 & 0 & 0 & 0 \\ -1 & 2 & 0 & 0 \\ 3 & -1 & 3 & 0 \\ 2 & 0 & -1 & 4 \end{bmatrix} \begin{bmatrix} 1 & -1 & 3 & 2 \\ 0 & 2 & -1 & 0 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 4 \end{bmatrix}$, $\begin{matrix} x_1 = 2 \\ x_2 = -3 \\ x_3 = 4 \\ x_4 = -1 \end{matrix}$

13. No, since $\mathbf{x}^\mathsf{T}(-\mathbf{A})\mathbf{x} = -\mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} < 0$; yes; yes; no

15. $\begin{bmatrix} -3.5 & 1.25 \\ 3.0 & -1.0 \end{bmatrix}$

17. $\frac{1}{36}\begin{bmatrix} 584 & 104 & -66 \\ 104 & 20 & -12 \\ -66 & -12 & 9 \end{bmatrix}$

19. $\frac{1}{16}\begin{bmatrix} 21 & -6 & -14 & 6 \\ -6 & 36 & -12 & -4 \\ -14 & -12 & 20 & -4 \\ 6 & -4 & -4 & 4 \end{bmatrix}$

Problem Set 20.3, page 863

5. Exact 0.5, 0.5, 0.5

7. $x_1 = 2$, $x_2 = -4$, $x_3 = 8$

9. Exact 2, 1, 4

11. (a) $\mathbf{x}^{(3)\mathsf{T}} = [0.49983 \quad 0.50001 \quad 0.500017]$,
(b) $\mathbf{x}^{(3)\mathsf{T}} = [0.50333 \quad 0.49985 \quad 0.49968]$

13. 8, −16, 43, 86 steps; spectral radius 0.09, 0.35, 0.72, 0.85, approximately

15. $[1.99934 \quad 1.00043 \quad 3.99684]^\mathsf{T}$ (Jacobi, Step 5); $[2.00004 \quad 0.998059 \quad 4.00072]^\mathsf{T}$ (Gauss–Seidel)

19. $\sqrt{306} = 17.49$, 12, 12

Problem Set 20.4, page 871

1. 18, $\sqrt{110} = 10.49$, 8, $[0.125 \quad -0.375 \quad 1 \quad 0 \quad -0.75 \quad 0]$

3. 5.9, $\sqrt{13.81} = 3.716$, 3, $\frac{1}{3}[0.2 \quad 0.6 \quad -2.1 \quad 3.0]$

5. 5, $\sqrt{5}$, 1, $[1 \quad 1 \quad 1 \quad 1 \quad 1]$

7. $ab + bc + ca = 0$

9. $\kappa = 5 \cdot \frac{1}{2} = 2.5$

11. $\kappa = (5 + \sqrt{5})(1 + 1/\sqrt{5}) = 6 + 2\sqrt{5}$

13. $\kappa = 19 \cdot 13 = 247$; ill-conditioned

15. $\kappa = 20 \cdot 20 = 400$; ill-conditioned

17. $167 \leqq 21 \cdot 15 = 315$

19. $[-2 \quad 4]^\mathsf{T}$, $[-144.0 \quad 184.0]^\mathsf{T}$, $\kappa = 25{,}921$, extremely ill-conditioned

21. Small residual $[0.145 \quad 0.120]$, but large deviation of $\tilde{\mathbf{x}}$.

23. 27, 748, 28,375, 943,656, 29,070,279

Problem Set 20.5, page 875

1. $1.846 - 1.038x$

3. $1.48 + 0.09x$

5. $s = 90t - 675$, $v_{\text{av}} = 90$ km/hr

9. $-11.36 + 5.45x - 0.589x^2$

11. $1.89 - 0.739x + 0.207x^2$

13. $2.552 + 16.23x$, $-4.114 + 13.73x + 2.500x^2$, $2.730 + 1.466x - 1.778x^2 + 2.852x^3$

Problem Set 20.7, page 884

1. 5, 0, 7; radii 6, 4, 6. Spectrum $\{-1, 4, 9\}$

3. Centers 0; radii 0.5, 0.7, 0.4. Skew-symmetric, hence $\lambda = i\mu$, $-0.7 \leqq \mu \leqq 0.7$.

5. 2, 3, 8; radii $1 + \sqrt{2}$, 1, $\sqrt{2}$; actually (4S) 1.163, 3.511, 8.326

7. $t_{11} = 100$, $t_{22} = t_{33} = 1$

9. They lie in the intervals with endpoints $a_{jj} \pm (n-1) \cdot 10^{-5}$. Why?

11. $\rho(\mathbf{A}) \leqq$ Row sum norm $\|\mathbf{A}\|_\infty = \max_j \sum_k |a_{jk}| = \max_j(|a_{jj}| + \text{Gerschgorin radius})$

13. $\sqrt{122} = 11.05$

15. $\sqrt{0.52} = 0.7211$

17. Show that $\mathbf{A}\overline{\mathbf{A}}^\mathsf{T} = \overline{\mathbf{A}}^\mathsf{T}\mathbf{A}$.

19. 0 lies in no Gerschgorin disk, by (3) with $>$; hence $\det \mathbf{A} = \lambda_1 \cdots \lambda_n \neq 0$.

Problem Set 20.8, page 887

1. $q = 10, 10.9908, 10.9999$; $|\epsilon| \leqq 3, 0.3028, 0.0275$

3. $q \pm \delta = 4 \pm 1.633$, 4.786 ± 0.619, 4.917 ± 0.398

5. Same answer as in Prob. 3, possibly except for small roundoff errors.

7. $q = 5.5, 5.5738, 5.6018$; $|\epsilon| \leqq 0.5, 0.3115, 0.1899$; eigenvalues (4S) 1.697, 3.382, 5.303, 5.618

9. $\mathbf{y} = \mathbf{A}\mathbf{x} = \lambda\mathbf{x}$, $\mathbf{y}^\mathsf{T}\mathbf{x} = \lambda\mathbf{x}^\mathsf{T}\mathbf{x}$, $\mathbf{y}^\mathsf{T}\mathbf{y} = \lambda^2\mathbf{x}^\mathsf{T}\mathbf{x}$,
$\epsilon^2 \leqq \mathbf{y}^\mathsf{T}\mathbf{y}/\mathbf{x}^\mathsf{T}\mathbf{x} - (\mathbf{y}^\mathsf{T}\mathbf{x}/\mathbf{x}^\mathsf{T}\mathbf{x})^2 = \lambda^2 - \lambda^2 = 0$

11. $q = 1, \cdots, -2.8993$ approximates -3 (0 of the given matrix), $|\epsilon| \leqq 1.633, \cdots, 0.7024$ (Step 8)

Problem Set 20.9, page 896

1. $$\begin{bmatrix} 0.98 & -0.4418 & 0 \\ -0.4418 & 0.8702 & 0.3718 \\ 0 & 0.3718 & 0.4898 \end{bmatrix}$$

3. $\begin{bmatrix} 7 & -3.6056 & 0 \\ -3.6056 & 13.462 & 3.6923 \\ 0 & 3.6923 & 3.5385 \end{bmatrix}$

5. $\begin{bmatrix} 3 & -67.59 & 0 & 0 \\ -67.59 & 143.5 & 45.35 & 0 \\ 0 & 45.35 & 23.34 & 3.126 \\ 0 & 0 & 3.126 & -33.87 \end{bmatrix}$

7. Eigenvalues 16, 6, 2

$$\begin{bmatrix} 11.2903 & -5.0173 & 0 \\ -5.0173 & 10.6144 & 0.7499 \\ 0 & 0.7499 & 2.0952 \end{bmatrix}, \begin{bmatrix} 14.9028 & -3.1265 & 0 \\ -3.1265 & 7.0883 & 0.1966 \\ 0 & 0.1966 & 2.0089 \end{bmatrix}, \begin{bmatrix} 15.8299 & -1.2932 & 0 \\ -1.2932 & 6.1692 & 0.0625 \\ 0 & 0.0625 & 2.0010 \end{bmatrix}$$

9. Eigenvalues (4S) 141.4, 68.64, −30.04

$$\begin{bmatrix} 141.1 & 4.926 & 0 \\ 4.926 & 68.97 & 0.8691 \\ 0 & 0.8691 & -30.03 \end{bmatrix}, \begin{bmatrix} 141.3 & 2.400 & 0 \\ 2.400 & 68.72 & 0.3797 \\ 0 & 0.3797 & -30.04 \end{bmatrix}, \begin{bmatrix} 141.4 & 1.166 & 0 \\ 1.166 & 68.66 & 0.1661 \\ 0 & 0.1661 & -30.04 \end{bmatrix}$$

Chapter 20 Review Questions and Problems, page 896

15. $[3.9 \quad 4.3 \quad 1.8]^T$

17. $[-2 \quad 0 \quad 5]^T$

19. $\begin{bmatrix} 0.28193 & -0.15904 & -0.00482 \\ -0.15904 & 0.12048 & -0.00241 \\ -0.00482 & -0.00241 & 0.01205 \end{bmatrix}$

21. $\begin{bmatrix} 5.750 \\ 3.600 \\ 0.838 \end{bmatrix}, \begin{bmatrix} 6.400 \\ 3.559 \\ 1.000 \end{bmatrix}, \begin{bmatrix} 6.390 \\ 3.600 \\ 0.997 \end{bmatrix}$

Exact: $[6.4 \quad 3.6 \quad 1.0]^T$

23. $\begin{bmatrix} 1.700 \\ 1.180 \\ 4.043 \end{bmatrix}, \begin{bmatrix} 1.986 \\ 0.999 \\ 4.002 \end{bmatrix}, \begin{bmatrix} 2.000 \\ 1.000 \\ 4.000 \end{bmatrix}$

Exact: $[2 \quad 1 \quad 4]^T$

25. 42, $\sqrt{674} = 25.96$, 21

27. 30

29. 5

31. $115 \cdot 0.4458 = 51.27$

33. $5 \cdot \frac{21}{63} = \frac{5}{3}$

35. $1.514 + 1.129x - 0.214x^2$

37. Centers 15, 35, 90; radii 30, 35, 25, respectively. Eigenvalues (3S) 2.63, 40.8, 96.6

39. Centers 0, −1, −4; radii 9, 6, 7, respectively; eigenvalues 0, 4.446, −9.446

Problem Set 21.1, page 910

1. $y = 5e^{-0.2x}$, 0.00458, 0.00830 (errors of y_5, y_{10})
3. $y = x - \tanh x$ (set $y - x = u$), 0.00929, 0.01885 (errors of y_5, y_{10})
5. $y = e^x$, 0.0013, 0.0042 (errors of y_5, y_{10})
7. $y = 1/(1 - x^2/2)$, 0.00029, 0.01187 (errors of y_5, y_{10})
9. Errors 0.03547 and 0.28715 of y_5 and y_{10} much larger
11. $y = 1/(1 - x^2/2)$; error -10^{-8}, $-4 \cdot 10^{-8}$, $\cdots$, $-6 \cdot 10^{-7}$, $+9 \cdot 10^{-6}$; $\epsilon = 0.0002/15 = 1.3 \cdot 10^{-5}$ (use RK with $h = 0.2$)
13. $y = \tan x$; error $0.83 \cdot 10^{-7}$, $0.16 \cdot 10^{-6}$, $\cdots$, $-0.56 \cdot 10^{-6}$, $+0.13 \cdot 10^{-5}$
15. $y = 3 \cos x - 2 \cos^2 x$; error $\cdot\ 10^7$: 0.18, 0.74, 1.73, 3.28, 5.59, 9.04, 14.3, 22.8, 36.8, 61.4
17. $y' = 1/(2 - x^4)$; error $\cdot\ 10^9$: 0.2, 3.1, 10.7, 23.2, 28.5, −32.3, −376, −1656, −3489, +80444
19. Errors for Euler–Cauchy 0.02002, 0.06286, 0.05074; for improved Euler–Cauchy −0.000455, 0.012086, 0.009601; for Runge–Kutta. 0.0000011, 0.000016, 0.000536

Problem Set 21.2, page 915

1. $y = e^x$, $y_5^* = 1.648717$, $y_5 = 1.648722$, $\epsilon_5 = -3.8 \cdot 10^{-8}$, $y_{10}^* = 2.718276$, $y_{10} = 2.718284$, $\epsilon_{10} = -1.8 \cdot 10^{-6}$
3. $y = \tan x$, $y_4, \cdots, y_{10}$ (error $\cdot\ 10^5$) 0.422798 (−0.49), 0.546315 (−1.2), 0.684161 (−2.4), 0.842332 (−4.4), 1.029714 (−7.5), 1.260288 (−13), 1.557626 (−22)
5. RK error smaller in absolute value, error $\cdot\ 10^5 = 0.4, 0.3, 0.2, 5.6$ (for $x = 0.4, 0.6, 0.8, 1.0$)
7. $y = 1/(4 + e^{-3x})$, $y_4, \cdots, y_{10}$ (error $\cdot\ 10^5$) 0.232490 (0.34), 0.236787 (0.44), 0.240075 (0.42), 0.242570 (0.35), 0.244453 (0.25), 0.245867 (0.16), 0.246926 (0.09)
9. $y = \exp(x^3) - 1$, $y_4, \cdots, y_{10}$ (error $\cdot\ 10^7$) 0.008032 (−4), 0.015749 (−10), 0.027370 (−17), 0.043810 (−26), 0.066096 (−39), 0.095411 (−54), 0.133156 (−74)
13. $y = \exp(x^2)$. Errors $\cdot\ 10^5$ from $x = 0.3$ to 0.7: −5, −11, −19, −31, −41
15. **(a)** 0, 0.02, 0.0884, 0.215848, $y_4 = 0.417818$, $y_5 = 0.708887$ (poor)
(b) By 30–50%

Problem Set 21.3, page 922

1. $y_1 = -e^{-2x} + 4e^x$, $y_2 = -e^{-2x} + e^x$; errors of y_1 (of y_2) from 0.002 to 0.5 (from −0.01 to 0.1), monotone
3. $y_1' = y_2$, $y_2' = -\frac{1}{4}y_1$, $y = y_1 = 1$, 0.99, 0.97, 0.94, 0.9005, error −0.005, −0.01, −0.015, −0.02, −0.0229; exact $y = \cos \frac{1}{2}x$
5. $y_1' = y_2$, $y_2' = y_1 + x$, $y_1(0) = 1$, $y_2(0) = -2$, $y = y_1 = e^{-x} - x$, $y = 0.8$ (error 0.005), 0.61 (0.01), 0.429 (0.012), 0.2561 (0.0142), 0.0905 (0.0160)
7. By about a factor 10^5. $\epsilon_n(y_1) \cdot 10^6 = -0.082, \cdots, -0.27$, $\epsilon_n(y_2) \cdot 10^6 = 0.08, \cdots, 0.27$
9. Errors of y_1 (of y_2) from $0.3 \cdot 10^{-5}$ to $1.3 \cdot 10^{-5}$ (from $0.3 \cdot 10^{-5}$ to $0.6 \cdot 10^{-5}$)
11. $(y_1, y_2) = (0, 1), (0.20, 0.98), (0.39, 0.92), \cdots, (-0.23, -0.97), (-0.42, -0.91), (-0.59), (-0.81)$; continuation will give an "ellipse."

Problem Set 21.4, page 930

3. $-3u_{11} + u_{12} = -200, \quad u_{11} - 3u_{12} = -100$
5. 105, 155, 105, 115; Step 5: 104.94, 154.97, 104.97, 114.98
7. 0, 0, 0, 0. All equipotential lines meet at the corners (why?).
Step 5: 0.29298, 0.14649, 0.14649, 0.073245
9. 0.108253, 0.108253, 0.324760, 0.324760; Step 10: 0.108538, 0.108396, 0.324902, 0.324831
11. **(a)** $u_{11} = -u_{12} = -66$. **(b)** Reduce to 4 equations by symmetry.
$u_{11} = u_{31} = -u_{15} = -u_{35} = -92.92$, $u_{21} = -u_{25} = -87.45$,
$u_{12} = u_{32} = -u_{14} = -u_{34} = -64.22$, $u_{22} = -u_{24} = -53.98$,
$u_{13} = u_{23} = u_{33} = 0$
13. $u_{12} = u_{32} = 31.25, \quad u_{21} = u_{23} = 18.75, \quad u_{jk} = 25$ at the others
15. $u_{21} = u_{23} = 0.25, \quad u_{12} = u_{32} = -0.25, \quad u_{jk} = 0$ otherwise
17. $\sqrt{3}$, $u_{11} = u_{21} = 0.0849$, $u_{12} = u_{22} = 0.3170$. (0.1083, 0.3248 are 4S-values of the solution of the linear system of the problem.)

Problem Set 21.5, page 935

5. $u_{11} = 0.766, \quad u_{21} = 1.109, \quad u_{12} = 1.957, u_{22} = 3.293$
7. $\mathbf{A}$, as in Example 1, right sides −220, −220, −220, −220.
Solution $u_{11} = u_{21} = 125.7$, $u_{21} = u_{22} = 157.1$
13. $-4u_{11} + u_{21} + u_{12} = -3$, $u_{11} - 4u_{21} + u_{22} = -12$, $u_{11} - 4u_{12} + u_{22} = 0$,
$2u_{21} + 2u_{12} - 12u_{22} = -14$, $u_{11} = u_{22} = 2$, $u_{21} = 4$, $u_{12} = 1$.
Here $-\frac{14}{3} = -\frac{4}{3}(1 + 2.5)$ with $\frac{4}{3}$ from the stencil.
15. $\mathbf{b} = [-200, -100, -100, 0]^\mathsf{T}; \quad u_{11} = 73.68$, $u_{21} = u_{12} = 47.37$, $u_{22} = 15.79$ (4S)

Problem Set 21.6, page 941

5. 0, 0.6625, 1.25, 1.7125, 2, 2.1, 2, 1.7125, 1.25, 0.6625, 0
7. Substantially less accurate, 0.15, 0.25 ($t = 0.04$), 0.100, 0.163 ($t = 0.08$)
9. Step 5 gives 0, 0.06279, 0.09336, 0.08364, 0.04707, 0.
11. Step 2: 0 (exact 0), 0.0453 (0.0422), 0.0672 (0.0658), 0.0671 (0.0628), 0.0394 (0.0373), 0 (0)
13. 0.3301, 0.5706, 0.4522, 0.2380 ($t = 0.04$), 0.06538, 0.10603, 0.10565, 0.6543 ($t = 0.20$)
15. 0.1018, 0.1673, 0.1673, 0.1018 ($t = 0.04$), 0.0219, 0.0355, $\cdots$ ($t = 0.20$)

Problem Set 21.7, page 944

1. $u(x, 1) = 0, -0.05, -0.10, -0.15, -0.20, 0$
3. For $x = 0.2, 0.4$ we obtain 0.24, 0.40 ($t = 0.2$), 0.08, 0.16 ($t = 0.4$), −0.08, −0.16 ($t = 0.6$), etc.
5. 0, 0.354, 0.766, 1.271, 1.679, 1.834, $\cdots$ ($t = 0.1$); 0, 0.575, 0.935, 1.135, 1.296, 1.357, $\cdots$ ($t = 0.2$)
7. 0.190, 0.308, 0.308, 0.190, (3S-exact: 0.178, 0.288, 0.288, 0.178)

Chapter 21 Review Questions and Problems, page 945

17. $y = e^x$, 0.038, 0.125 (errors of y_5 and y_{10})
19. $y = \tan x$; 0 (0), 0.10050 (−0.00017), 0.20304 (−0.00033), 0.30981 (−0.00048), 0.42341 (−0.00062), 0.54702 (−0.00072), 0.68490 (−0.00076), 0.84295 (−0.00066), 1.0299 (−0.0002), 1.2593 (0.0009), 1.5538 (0.0036)
21. 0.1003346 $(0.8 \cdot 10^{-7})$ 0.2027099 $(1.6 \cdot 10^{-7})$, 0.3093360 $(2.1 \cdot 10^{-7})$, 0.4227930 $(2.3 \cdot 10^{-7})$, 0.5463023 $(1.8 \cdot 10^{-7})$
23. $y = \sin x$, $y_{0.8} = 0.717366$, $y_{1.0} = 0.841496$ (errors $-1.0 \cdot 10^{-5}$, $-2.5 \cdot 10^{-5}$)
25. $y_1' = y_2$, $y_2' = x^2 y_1$, $y = y_1 = 1, 1, 1, 1.0001, 1.0006, 1.002$
27. $y_1' = y_2$, $y_2' = 2e^x - y_1$, $y = e^x - \cos x$, $y = y_1 = 0, 0.241, 0.571, \cdots$; errors between 10^{-6} and 10^{-5}
29. 3.93, 15.71, 58.93
31. 0, 0.04, 0.08, 0.12, 0.15, 0.16, 0.15, 0.12, 0.08, 0.04, 0 ($t = 0.3$. 3 time steps)
33. $u(P_{11}) = u(P_{31}) = 270$, $u(P_{21}) = u(P_{13}) = u(P_{23}) = u(P_{33}) = 30$, $u(P_{12}) = u(P_{32}) = 90$, $u(P_{22}) = 60$
35. 0.043330, 0.077321, 0.089952, 0.058488 ($t = 0.04$), 0.010956, 0.017720, 0.017747, 0.010964 ($t = 0.20$)

Problem Set 22.1, page 953

3. $f(\mathbf{x}) = 2(x_1 - 1)^2 + (x_2 + 2)^2 - 6$; Step 3: (1.037, −1.926), value −5.992
9. Step 5: (0.11247, −0.00012), value 0.000016

Problem Set 22.2, page 957

7. No
9. x_3, x_4 is the unused time on M_1, M_2, respectively.
11. $f(2.5, 2.5) = 100$
13. $f(-\frac{11}{3}, \frac{26}{3}) = 198\frac{1}{3}$
15. $f(9, 6) = 360$
17. $0.5x_1 + 0.75x_2 \leqq 45$ (copper), $0.5x_1 + 0.25x_2 \leqq 30$, $f = 120x_1 + 100x_2$, $f_{\max} = f(45, 30) = 8400$
19. $f = x_1 + x_2$, $2x_1 + 3x_2 \leqq 1200$, $4x_1 + 2x_2 \leqq 1600$, $f_{\max} = f(300, 200) = 500$
21. $x_1/3 + x_2/2 \leqq 100$, $x_1/3 + x_2/6 \leqq 80$, $f = 150x_1 + 100x_2$, $f_{\max} = f(210, 60) =$ 37,500

Problem Set 22.3, page 961

3. $f(120/11, 60/11) = 480/11$
5. Eliminate in Column 3, so that 20 goes. $f_{\min} = f(0, \frac{1}{2}) = -10$.
7. $f_{\max} = f(\frac{60}{21}, 0, \frac{1500}{105}, 0) = \frac{2200}{7}$
9. $f_{\max} = 6$ on the segment from (3, 0, 0) to (0, 0, 2)
11. We minimize! The augmented matrix is

$$\mathbf{T}_0 = \begin{bmatrix} 1 & 1.8 & 2.1 & 0 & 0 & 0 \\ 0 & 15 & 30 & 1 & 0 & 150 \\ 0 & 600 & 500 & 0 & 1 & 3900 \end{bmatrix}.$$

The pivot is 600. The calculation gives

$$\mathbf{T}_1 = \begin{bmatrix} 1 & 0 & \frac{6}{10} & 0 & -\frac{3}{1000} & -\frac{117}{10} \\ 0 & 0 & \frac{35}{2} & 1 & -\frac{1}{40} & \frac{105}{2} \\ 0 & 600 & 500 & 0 & 1 & 3900 \end{bmatrix} \quad \begin{matrix} \text{Row 1} - \frac{1.8}{600}\text{Row 3} \\ \text{Row 2} - \frac{15}{600}\text{Row 3} \\ \text{Row 3} \end{matrix}$$

The next pivot is $\frac{35}{2}$. The calculation gives

$$\mathbf{T}_2 = \begin{bmatrix} 1 & 0 & 0 & -\frac{6}{175} & -\frac{3}{1400} & -\frac{27}{2} \\ 0 & 0 & \frac{35}{2} & 1 & -\frac{1}{40} & \frac{105}{2} \\ 0 & 600 & 0 & -\frac{200}{7} & \frac{12}{7} & 2400 \end{bmatrix} \quad \begin{matrix} \text{Row 1} - \frac{1.2}{35}\text{Row 2} \\ \text{Row 2} \\ \text{Row 3} - \frac{1000}{35}\text{Row 2} \end{matrix}$$

Hence $-f$ has the maximum value -13.5, so that f has the minimum value 13.5, at the point

$$(x_1, x_2) = \left(\frac{2400}{600}, \frac{105/2}{35/2}\right) = (4, 3).$$

13. $f_{\max} = f(5, 4, 6) = 478$

Problem Set 22.4, page 968

1. $f(6, 3) = 84$
3. $f(20, 20) = 40$
5. $f(10, 5) = 5500$
7. $f(1, 1, 0) = 13$
9. $f(4, 0, \frac{1}{2}) = 9$

Chapter 22 Review Questions and Problems, page 968

9. Step 5: $[0.353 \quad -0.028]^{\mathsf{T}}$. Slower. Why?
11. Of course! Step 5: $[-1.003 \quad 1.897]^{\mathsf{T}}$
17. $f(2, 4) = 100$
19. $f(3, 6) = -54$

Problem Set 23.1, page 974

9. $\begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$

11. $\begin{bmatrix} 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$

13. $\begin{bmatrix} 0 & 1 & 1 \\ 0 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$

15.

17. If G is complete.

19.

$$\text{Vertex}\quad\begin{array}{c|c} & \text{Edge} \\ & \begin{array}{cccc} e_1 & e_2 & e_3 & e_4 \end{array} \\ \hline \begin{array}{c} 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{bmatrix} -1 & -1 & 1 & -1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \end{array}$$

Problem Set 23.2, page 979

1. 5

3. 4

5. The idea is to go backward. There is a v_{k-1} adjacent to v_k and labeled $k - 1$, etc. Now the only vertex labeled 0 is s. Hence $\lambda(v_0) = 0$ implies $v_0 = s$, so that $v_0 - v_1 - \cdots - v_{k-1} - v_k$ is a path $s \rightarrow v_k$ that has length k.

15. Delete the edge (2, 4).

17. No

Problem Set 23.3, page 983

1. (1, 2), (2, 4), (4, 3); $L_2 = 12, L_3 = 36, L_4 = 28$

5. (1, 2), (2, 4), (3, 4), (3, 5); $L_2 = 2, L_3 = 4, L_4 = 3, L_5 = 6$

7. (1, 2), (2, 4), (3, 4); $L_2 = 10, L_3 = 15, L_4 = 13$

9. (1, 5), (2, 3), (2, 6), (3, 4), (3, 5); $L_2 = 9, L_3 = 7, L_4 = 8, L_5 = 4, L_6 = 14$

Problem Set 23.4, page 987

1.
```
2\
  >4 — 3 — 5    L = 10
1/
```

3.
```
           1
          /
5 — 3 — 6          L = 17
          \
           2 — 4
```

5.
```
   2
  /
1       3       L = 12
  \   /
    4
      \
        5
```

9. Yes

11.
```
           2
          /
1 — 3 — 4          L = 38
          \
           5 — 6
```

13. New York–Washington–Chicago–Dalles–Denver–Los Angeles

15. G is connected. If G were not a tree, it would have a cycle, but this cycle would provide two paths between any pair of its vertices, contradicting the uniqueness.

19. If we add an edge (u, v) to T, then since T is connected, there is a path $u \to v$ in T which, together with (u, v), forms a cycle.

Problem Set 23.5, page 990

1. If G is a tree.
3. A shortest spanning tree of the largest connected graph that contains vertex 1.
7. (1, 4), (1, 3), (1, 2), (2, 6), (3, 5); $L = 32$
9. (1, 4), (4, 3), (4, 2), (3, 5); $L = 20$
11. (1, 4), (4, 3), (4, 5), (1, 2); $L = 12$

Problem Set 23.6, page 997

1. $\{3, 6\}$, $11 + 3 = 14$
3. $\{4, 5, 6\}$, $10 + 5 + 13 = 28$
5. $\{3, 6, 7\}$, $8 + 4 + 4 = 16$
7. $S = \{1, 4\}$, $8 + 6 = 14$
9. One is interested in flows *from s to t,* not in the opposite direction.
13. $\Delta_{12} = 5, \Delta_{24} = 8, \Delta_{45} = 2;$ $\Delta_{12} = 5, \Delta_{25} = 3;$ $\Delta_{13} = 4, \Delta_{35} = 9$
P_1: $1 - 2 - 4 - 5, \Delta f = 2;$ P_2: $1 - 2 - 5, \Delta f = 3;$ P_3: $1 - 3 - 5, \Delta f = 4$
15. $1 - 2 - 5, \Delta f = 2;$ $1 - 4 - 2 - 5, \Delta f = 2$, etc.
17. $f_{13} = f_{35} = 8,$ $f_{14} = f_{45} = 5,$ $f_{12} = f_{24} = f_{46} = 4,$ $f_{56} = 13,$ $f = 4 + 13 = 17,$ $f = 17$ is unique.
19. For instance, $f_{12} = 10,$ $f_{24} = f_{45} = 7,$ $f_{13} = f_{25} = 5,$ $f_{35} = 3,$ $f_{32} = 2,$ $f = 3 + 5 + 7 = 15,$ $f = 15$ is unique.

Problem Set 23.7, page 1000

3. (2, 3) and (5, 6)
5. By considering only edges with one labeled end and one unlabeled end
7. $1 - 2 - 5, \Delta_t = 2;$ $1 - 4 - 2 - 5, \Delta_t = 1;$ $f = 6 + 2 + 1 = 9$, where 6 is the given flow
9. $1 - 2 - 4 - 6, \Delta_t = 2;$ $1 - 3 - 5 - 6, \Delta_t = 1;$ $f = 4 + 2 + 1 = 7$, where 4 is the given flow
15. $S = \{1, 2, 4, 5\},$ $T = \{3, 6\},$ $\text{cap}(S, T) = 14$

Problem Set 23.8, page 1005

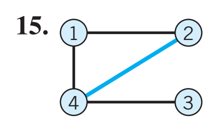
1. No **3.** No
5. Yes, $S = \{1, 4, 5, 8\}$
7. Yes, $S = \{1, 3, 5\}$ **11.** $1 - 2 \boldsymbol{-} 3 - 7 \boldsymbol{-} 5 - 4$
13. $1 - 2 \boldsymbol{-} 3 - 7 \boldsymbol{-} 5 - 4$ is augmenting and gives $1 \boldsymbol{-} 2 - 3 \boldsymbol{-} 7 - 5 \boldsymbol{-} 4$ and (1, 2), (3, 7), (5, 4) is of maximum cardinality.
15. $1 - 4 \boldsymbol{-} 3 - 6 \boldsymbol{-} 7 - 8$ is augmenting and gives $1 \boldsymbol{-} 4 - 3 \boldsymbol{-} 6 - 7 \boldsymbol{-} 8$ and (1, 4), (3, 6), (7, 8) is of maximum cardinality.
19. 3 **21.** 2
23. 3 **25.** K_4

Chapter 23 Review Questions and Problems, page 1006

11. $$\begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}$$

13. To vertex 1 2 3 4

From vertex
$$\begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix} \begin{bmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix}$$

15. 

17.

Vertex	Incident Edges
1	(1, 2), (1, 4)
2	(2, 1), (2, 4)
3	(3, 4)
4	(4, 1), (4, 2), (4, 3)

19. (1, 2), (1, 4), (2, 3); $L_2 = 2, L_3 = 5, L_4 = 5$

23. (1, 6), (4, 5), (2, 3), (7, 8)

Problem Set 24.1, page 1015

1. $q_L = 19, q_M = 20, q_U = 20.5$
3. $q_L = 138, q_M = 144, q_U = 154$
5. $q_L = 199, q_M = 201, q_U = 201$
7. $q_L = 1.3, q_M = 1.4, q_U = 1.45$
9. $q_L = 89.9, q_M = 91.0, q_U = 91.8$
11. $\bar{x} = 19.875, s = 0.835$, IQR $= 1.5$
13. $\bar{x} = 144.67, s = 8.9735$, IQR $= 16$
15. $\bar{x} = 1.355, s = 0.136$, IQR $= 0.15$
17. 3.54, 1.29

Problem Set 24.2, page 1017

1. 2^3 outcomes: *RRR, RRL, RLR, LRR, RLL, LRL, LLR, LLL*
3. $6^2 = 36$ outcomes (1, 1), (1, 2), $\cdots$, (6, 6), first number (second number) referring to the first die (second die)
5. Infinitely many outcomes *H TH TTH TTTH* $\cdots$ (*H* = *Head*, *T* = *Tail*)
7. The space of ordered pairs of numbers
9. 10 outcomes: *D ND NND* $\cdots$ *NNNNNNNNND*
11. Yes
17. $A \cup B = B$ implies $A \subseteq B$ by the definition of union. Conversely. $A \subseteq B$ implies that $A \cup B = B$ because always $B \subseteq A \cup B$, and if $A \subseteq B$, we must have equality in the previous relation.

Problem Set 24.3, page 1024

1. $1 - 4/216 = 98.15\%$, by Theorem 1
3. **(a)** $0.9^3 = 72.9\%$, **(b)** $\frac{90}{100} \cdot \frac{89}{99} \cdot \frac{88}{98} = 72.65\%$
5. $\frac{8}{9}$
7. Small sample from a large population containing *many* items in each class we are interested in (defectives and nondefectives, etc.)
9. $\frac{498}{500} \cdot \frac{497}{499} \cdot \frac{496}{498} \cdot \frac{495}{497} \cdot \frac{494}{496} \approx 0.98008$
11. **(a)** $\frac{100}{200} \cdot \frac{99}{199} = 24.874\%$, **(b)** $\frac{100}{200} \cdot \frac{100}{199} + \frac{100}{200} \cdot \frac{100}{199} = 50.25\%$, **(c)** same as (a). (a) + (b) + (c) = 1. Why?
13. $1 - 0.96^3 = 11.5\%$
15. $1 - 0.875^4 = 0.4138 < 1 - 0.75^2 = 0.4375 < 0.5 \quad (c < b < a)$
17. $A = B \cup (A \cap B^c)$, hence $P(A) = P(B) + P(A \cap B^c) \geqq P(B)$ by disjointedness of B and $A \cap B^c$

Problem Set 24.4, page 1028

1. In $10! = 3{,}628{,}800$ ways
3. $\frac{2}{6} \cdot \frac{1}{5} \cdot \frac{4}{4} \cdot \frac{3}{3} \cdot \frac{2}{2} \cdot \frac{1}{1} = \frac{4}{6} \cdot \frac{3}{5} \cdot \frac{2}{4} \cdot \frac{1}{3} \cdot \frac{2}{2} \cdot \frac{1}{1} = \frac{4!2!}{6!} = \frac{2}{6} \cdot \frac{1}{5} = \frac{1}{15}$
5. $\binom{10}{3}\binom{5}{2}\binom{6}{2} = 18{,}000$
7. 210, 70, 112, 28
9. In $6!/6 = 120$ ways
11. $9 \cdot 8 = 72$
13. **(b)** $1/(12n)$
15. P (*No two people have a birthday in common*) $= 365 \cdot 364 \cdots 346/365^{20} = 0.59$. *Answer:* 41%, which is surprisingly large.

Problem Set 24.5, page 1034

1. $k = \frac{1}{55}$ by (6)
3. $k = \frac{1}{4}$ by (10), $P(0 \leqq X \leqq 2) = \frac{1}{2}$
5. No, because of (6)
7. $k = \frac{1}{100}$ because of (6) and $1 + 8 + 27 + 64 = 100$
9. $k = 5$; 50%
11. $0.5^3 = 12.5\%$
13. $F(x) = 0$ if $x < -1$, $F(x) = \frac{1}{2}(x + 1)^2$ if $-1 \leqq x < 0$
$F(x) = 1 - \frac{1}{2}(x - 1)^2$ if $0 \leqq x < 1$, $F(x) = 1$ if $x \leqq 1$
Answer: 500 cans, $P = 0.125, 0$
15. $X > b, X \geqq b, X < c, X \leqq c$, etc.

Problem Set 24.6, page 1038

1. $k = \frac{1}{2}, \mu = \frac{4}{3}, \sigma^2 = \frac{2}{9}$
3. $\mu = \pi, \sigma^2 = \pi^2/3$; cf. Example 2
5. $\mu = \frac{1}{4}, \sigma^2 = \frac{1}{16}$
7. $C = \frac{1}{2}, \mu = 2, \sigma^2 = 4$
9. 750, 1, 0.002
11. $c = 0.073$
13. $643.50
15. $\frac{1}{2}, \frac{1}{20}, (X - \frac{1}{2})\sqrt{20}$
17. X = *Product of the 2 numbers*. $E(X) = 12.25$, 12 cents
19. $(0 + 1 \cdot 3 + 3 \cdot 8 + 1 \cdot 27)/8 = 54/8 = 6 \cdot 75$

Problem Set 24.7, page 1044

3. 38%
5. $\binom{5}{x}\,0.5^5$, 0.03125, 0.15625, $1 - f(0) = 0.96875$, 0.96875
7. 0.265
9. $f(x) = 0.5^x e^{-0.5}/x!$, $f(0) + f(1) = e^{-0.5}(1.0 + 0.5) = 0.91$. *Answer:* 9%
11. $13\frac{1}{4}$ %
13. 42%, 47.2%, 10.5%, 0.3%
15. $1 - e^{-0.2} = 18\%$

Problem Set 24.8, page 1050

1. 0.1587, 0.5, 0.6915, 0.6247
3. 45.065, 56.978, 2.022
5. 15.9%
7. 31.1%, 95.4%
9. About 58%
11. $t = 1084$ hours
13. About 683 (Fig. 521a)

Problem Set 24.9, page 1059

1. $\frac{1}{8}, \frac{3}{16}, \frac{3}{8}$
3. $\frac{2}{9}, \frac{1}{9}, \frac{1}{2}$
5. $f_2(y) = 1/(\beta_2 - \alpha_2)$ if $\alpha_2 < y < \beta_2$
7. 27.45 mm, 0.38 mm
11. 25.26 cm, 0.0078 cm
13. 50%
15. The distributions in Prob. 17 and Example 1
17. No

Chapter 24 Review Questions and Problems, page 1060

11. $Q_L = 110$, $Q_M = 112$, $Q_U = 115$
13. $\bar{x} = 111.9$, $s = 4.0125$, $s^2 = 16.1$
21. $x_{\text{min}} \leqq x_j \leqq x_{\text{max}}$. Sum over j from 1.
17. $\bar{x} = 6$, $s = 3.65$
19. $f(x) = \binom{50}{x} 0.03^x 0.97^{50-x} \approx 1.5^x e^{-1.5}/x!$
21. $f(x) = 2^{-x}$, $x = 1, 2, \cdots$
23. $1, \frac{1}{2}$
25. 0.1587, 0.6306, 0.5, 0.4950

Problem Set 25.2, page 1067

1. In Example 1, $\mu = 0$ so $\sum_{j=1}^{n} x_j = 0$. $\partial \ln \ell/\partial \ell = 0$ and $\tilde{\sigma}^2$ is as before.
3. $\ell = e^{-n\mu}\mu^{(x_1+\cdots+x_n)}/(x_1!\cdots x_n!)$, $\partial \ln \ell/\partial\mu = -n + (x_1 + \cdots + x_n)/\mu = 0$, $n\hat{\mu} = n\bar{x}$, $\hat{\mu} = \bar{x} = 15.3$
5. $l = p^k(1-p)^{n-k}$, $\hat{p} = k/n$, $k =$ number of successes in n trails
7. 7/12
9. $l = f = p(1-p)^{x-1}$, etc., $\hat{p} = 1/x$
11. $\hat{\theta} = n/\Sigma x_j = 1/\bar{x}$
13. $\hat{\theta} = 1$
15. Variability larger than perhaps expected

Problem Set 25.3, page 1077

3. Shorter by a factor $\sqrt{2}$

5. 4, 16

7. $c = 1.96, \bar{x} = 126, s^2 = 126 \cdot 674/800 = 106.155, k = cs/\sqrt{n} = 0.714$, $\text{CONF}_{0.95}\{125.3 \leqq \mu \leqq 126.7\}$, $\text{CONF}_{0.95}\{0.1566 \leqq p \leqq 0.1583\}$

9. $\text{CONF}_{0.99}\{63.72 \leqq \mu \leqq 66.28\}$

11. $n - 1 = 5, F(c) = 0.995, c = 4.03, \bar{x} = 9533.33, s^2 = 49{,}666.67$, $k = 366.66$ (Table 25.2), $\text{CONF}_{0.99}\{9166.7 \leqq \mu \leqq 9900\}$

13. $\text{CONF}_{0.95}\{0.023 \leqq \sigma^2 \leqq 0.085\}$

15. $n - 1 = 99$ degrees of freedom. $F(c_1) = 0.025$, $c_1 = 74.2$, $F(c_2) = 0.975$, $c_2 = 129.6$. Hence $k_1 = 12.41$, $k_2 = 7.10$. $\text{CONF}_{0.95}\{7.10 \leqq \sigma^2 \leqq 12.41\}$.

17. $\text{CONF}_{0.95}\{0.74 \leqq \sigma^2 \leqq 5.19\}$

19. $Z = X + Y$ is normal with mean 105 and variance 1.25. *Answer:* $P(104 \leqq Z \leqq 106) = 63\%$

Problem Set 25.4, page 1086

3. $t = (0.286 - 0)/(4.31/\sqrt{7}) = 0.18 < c = 1.94$; accept the hypothesis.

5. $c = 6090 > 6019$: do not reject the hypothesis.

7. $\sigma^2/n = 1.8$, $c = 57.8$, accept the hypothesis.

9. $\mu < 58.69$ or $\mu > 61.31$

11. Alternative $\mu \neq 5000$, $t = (4990 - 5000)/(20/\sqrt{50}) = -3.54 < c = -2.01$ (Table A9, Appendix 5). Reject the hypothesis $\mu = 5000$ g.

13. Two-sided. $t = (0.55 - 0)/\sqrt{0.546/8} = 2.11 < c = 2.37$ (Table A9, Appendix 5), no difference

15. $19 \cdot 1.0^2/0.8^2 = 29.69 < c = 30.14$ (Table A10. Appendix 5), accept the hypothesis

17. By (12), $t_0 = \sqrt{16}(20.2 - 19.6)/\sqrt{0.16 + 0.36} > c = 1.70$. Assert that B is better.

Problem Set 25.5, page 1091

1. $\text{LCL} = 1 - 2.58 \cdot 0.02/2 = 0.974$, $\text{UCL} = 1.026$

3. 27

5. Choose 4 times the original sample size

9. $2.58\sqrt{0.0004}/\sqrt{2} = 0.036$, $\text{LCL} = 3.464$, $\text{UCL} = 3.536$

11. $\text{LCL} = np - 3\sqrt{np(1-p)}$, $\text{CL} = np$, $\text{UCL} = np + 3\sqrt{np(1-p)}$

13. In about 30% (5%) of the cases

15. $\text{LCL} = \mu - 3\sqrt{\mu}$ is negative in (b) and we set $\text{LCL} = 0$, $\text{CL} = \mu = 3.6$, $\text{UCL} = \mu + 3\sqrt{\mu} = 9.3$.

Problem Set 25.6, page 1095

1. 0.9825, 0.9384, 0.4060

3. 0.8187, 0.6703, 0.1353

5. $e^{-25\theta}(1 + 25\theta)$, $P(A; 1.5) = 94.5$, $\alpha = 5.5\%$

7. 19.5%, 14.7%

9. $(1 - \theta)^n + n\theta(1 - \theta)^{n-1}$

11. $(1 - \frac{1}{2})^3 + 3 \cdot \frac{1}{2}(1 - \frac{1}{2})^2 = \frac{1}{2}$

13. $\sum_{x=0}^{9} \binom{100}{x} 0.12^x\, 0.88^{100-x} = 22\%$ (by the normal approximation)

15. $(1 - \theta)^5$, $[\theta(1 - \theta)^{5-1}]' = 0$, $\theta = \frac{1}{6}$, AOQL = 6.7%

Problem Set 25.7, page 1099

3. $\chi_0^2 = (40 - 50)^2/50 + (60 - 50)^2/50 = 4 > c = 3.84$; no
5. $\chi_0^2 = \frac{16}{10} > 11.07$; yes
7. $\chi_0^2 = 10.264 < 11.07$; yes
9. 42 even digits, accept.
13. $\chi_0^2 = \dfrac{(355 - 358.5)^2}{358.5} + \dfrac{(123 - 119.5)^2}{119.5} = 0.137 < c = 3.84$ (1 degree of freedom, 95%)
15. Combining the last three nonzero values, we have $K - r - 1 = 9$ ($r = 1$ since we estimated the mean, $\frac{10{,}094}{2608} \approx 3.87$). $\chi_0^2 = 12.8 < c = 16.92$. Accept the hypothesis.

Problem Set 25.8, page 1102

3. $(\frac{1}{2})^8 + 8 \cdot (\frac{1}{2})^8 = 3.5\%$ is the probability that 7 cases in 8 trials favor A under the hypothesis that A and B are equally good. Reject.
5. $(\frac{1}{2})^{18}(1 + 18 + 153 + 816) = 0.0038$
7. $\bar{x} = 9.67$, $s = 11.87$. $t_0 = 9.67/(11.87/\sqrt{15}) = 3.16 > c = 1.76$ ($\alpha = 5\%$). Hypothesis rejected.
9. Hypothesis $\tilde{\mu} = 0$. Alternative $\tilde{\mu} > 0$, $\bar{x} = 1.58$, $t = \sqrt{10} \cdot 1.58/1.23 = 4.06 > c = 1.83$ ($\alpha = 5\%$). Hypothesis rejected.
11. Consider $y_j = x_j - \tilde{\mu}_0$.
13. $n = 8$; 4 transpositions, $P(T \leqq 4) = 0.007$. Assert that fertilizing increases yield.
15. $P(T \leqq 2) = 2.8\%$. Assert that there is an increase.

Problem Set 25.9, page 1111

1. $y = 0.98 + 0.495x$
3. $y = -11{,}457.9 + 43.2x$
5. $y = -10 + 0.55x$
7. $y = 0.5932 + 0.1138x$, $R = 1/0.1138$
9. $y = 0.32923 + 0.00032x$, $y(66) = 0.35035$
13. $c = 3.18$ (Table A9), $k_1 = 43.2$, $q_0 = 54{,}878$, $K = 1.502$, $\text{CONF}_{0.95}\{41.7 \leqq \kappa_1 \leqq 44.7\}$.
15. $y - 1.875 = 0.067(x - 25)$, $3s_x^2 = 500$, $q_0 = 0.023$, $K = 0.021$, $\text{CONF}_{0.95}\{0.046 \leqq \kappa_1 \leqq 0.088\}$

Chapter 25 Review Questions and Problems, page 1111

15. $\hat{\mu} = 20.325$, $\hat{\sigma}^2 = (\frac{7}{8})s^2 = 3.982$
17. $\text{CONF}_{0.99}\{27.94 \leqq \mu \leqq 34.81\}$
19. $c = 14.74 > 14.5$, reject μ_0; $\Phi((14.74 - 14.50)/\sqrt{0.025}) = 0.9353$
21. $2.58 \cdot \sqrt{0.00024}/\sqrt{2} = 0.028$, LCL $= 2.722$, UCL $= 2.778$
23. $\alpha = 1 - (1 - \theta)^6 = 5.85\%$, when $\theta = 0.01$. For $\theta = 15\%$ we obtain $\beta = (1 - \theta)^6 = 37.7\%$. If n increases, so does α, whereas β decreases.
25. $y = 3.4 - 1.85x$

APPENDIX 3

Auxiliary Material

A3.1 Formulas for Special Functions

For tables of numeric values, see Appendix 5.

Exponential function e^x (Fig. 545)

$$e = 2.71828\ 18284\ 59045\ 23536\ 02874\ 71353$$

(1) $$e^x e^y = e^{x+y}, \qquad e^x/e^y = e^{x-y}, \qquad (e^x)^y = e^{xy}$$

Natural logarithm (Fig. 546)

(2) $$\ln(xy) = \ln x + \ln y, \qquad \ln(x/y) = \ln x - \ln y, \qquad \ln(x^a) = a \ln x$$

$\ln x$ is the inverse of e^x, and $e^{\ln x} = x$, $e^{-\ln x} = e^{\ln(1/x)} = 1/x$.

Logarithm of base ten $\log_{10} x$ or simply $\log x$

(3) $$\log x = M \ln x, \qquad M = \log e = 0.43429\ 44819\ 03251\ 82765\ 11289\ 18917$$

(4) $$\ln x = \frac{1}{M} \log x, \qquad \frac{1}{M} = \ln 10 = 2.30258\ 50929\ 94045\ 68401\ 79914\ 54684$$

$\log x$ is the inverse of 10^x, and $10^{\log x} = x$, $10^{-\log x} = 1/x$.

Sine and cosine functions (Figs. 547, 548). In calculus, angles are measured in radians, so that $\sin x$ and $\cos x$ have period 2π.

$\sin x$ is odd, $\sin(-x) = -\sin x$, and $\cos x$ is even, $\cos(-x) = \cos x$.

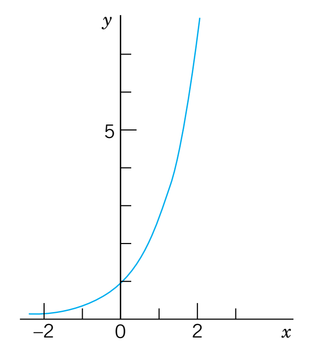


Fig. 545. Exponential function e^x

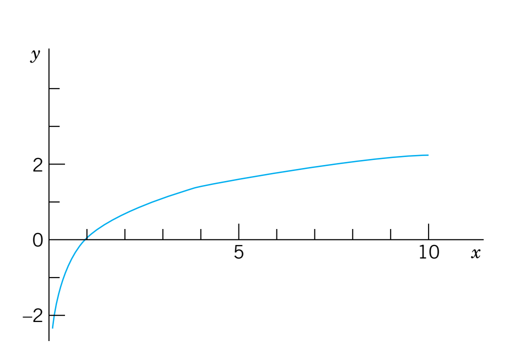


Fig. 546. Natural logarithm $\ln x$

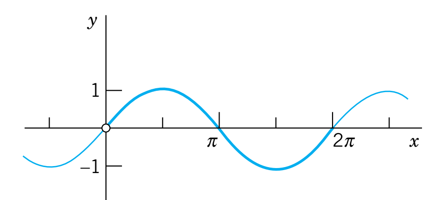


Fig. 547. sin x

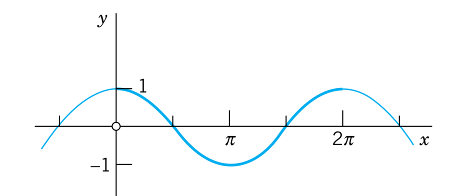


Fig. 548. cos x

$$1^\circ = 0.01745\ 32925\ 19943 \text{ radian}$$

$$1 \text{ radian} = 57^\circ\ 17'\ 44.80625''$$

$$= 57.29577\ 95131^\circ$$

$$\sin^2 x + \cos^2 x = 1 \tag{5}$$

$$\begin{cases} \sin(x+y) = \sin x \cos y + \cos x \sin y \\ \sin(x-y) = \sin x \cos y - \cos x \sin y \\ \cos(x+y) = \cos x \cos y - \sin x \sin y \\ \cos(x-y) = \cos x \cos y + \sin x \sin y \end{cases} \tag{6}$$

$$\sin 2x = 2 \sin x \cos x, \qquad \cos 2x = \cos^2 x - \sin^2 x \tag{7}$$

$$\begin{cases} \sin x = \cos\left(x - \dfrac{\pi}{2}\right) = \cos\left(\dfrac{\pi}{2} - x\right) \\ \cos x = \sin\left(x + \dfrac{\pi}{2}\right) = \sin\left(\dfrac{\pi}{2} - x\right) \end{cases} \tag{8}$$

$$\sin(\pi - x) = \sin x, \qquad \cos(\pi - x) = -\cos x \tag{9}$$

$$\cos^2 x = \tfrac{1}{2}(1 + \cos 2x), \qquad \sin^2 x = \tfrac{1}{2}(1 - \cos 2x) \tag{10}$$

$$\begin{cases} \sin x \sin y = \tfrac{1}{2}[-\cos(x+y) + \cos(x-y)] \\ \cos x \cos y = \tfrac{1}{2}[\cos(x+y) + \cos(x-y)] \\ \sin x \cos y = \tfrac{1}{2}[\sin(x+y) + \sin(x-y)] \end{cases} \tag{11}$$

$$\begin{cases} \sin u + \sin v = 2 \sin \dfrac{u+v}{2} \cos \dfrac{u-v}{2} \\ \cos u + \cos v = 2 \cos \dfrac{u+v}{2} \cos \dfrac{u-v}{2} \\ \cos v - \cos u = 2 \sin \dfrac{u+v}{2} \sin \dfrac{u-v}{2} \end{cases} \tag{12}$$

$$A \cos x + B \sin x = \sqrt{A^2 + B^2} \cos(x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \mp \frac{B}{A} \tag{13}$$

$$A \cos x + B \sin x = \sqrt{A^2 + B^2} \sin(x \pm \delta), \qquad \tan \delta = \frac{\sin \delta}{\cos \delta} = \pm \frac{A}{B} \tag{14}$$

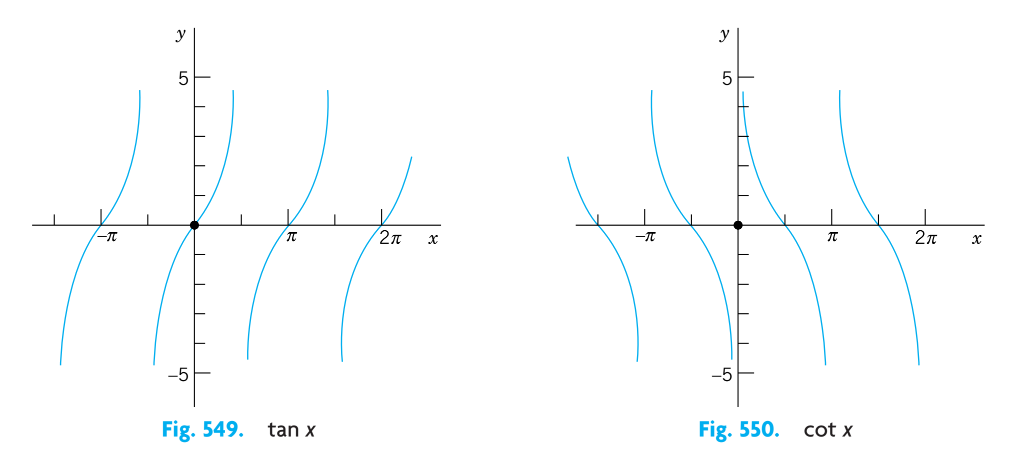


Fig. 549. tan x **Fig. 550.** cot x

Tangent, cotangent, secant, cosecant (Figs. 549, 550)

$$\tan x = \frac{\sin x}{\cos x}, \qquad \cot x = \frac{\cos x}{\sin x}, \qquad \sec x = \frac{1}{\cos x}, \qquad \csc x = \frac{1}{\sin x} \tag{15}$$

$$\tan(x + y) = \frac{\tan x + \tan y}{1 - \tan x \tan y}, \qquad \tan(x - y) = \frac{\tan x - \tan y}{1 + \tan x \tan y} \tag{16}$$

Hyperbolic functions (hyperbolic sine sinh x, etc.; Figs. 551, 552)

$$\sinh x = \tfrac{1}{2}(e^x - e^{-x}), \qquad \cosh x = \tfrac{1}{2}(e^x + e^{-x}) \tag{17}$$

$$\tanh x = \frac{\sinh x}{\cosh x}, \qquad \coth x = \frac{\cosh x}{\sinh x} \tag{18}$$

$$\cosh x + \sinh x = e^x, \qquad \cosh x - \sinh x = e^{-x} \tag{19}$$

$$\cosh^2 x - \sinh^2 x = 1 \tag{20}$$

$$\sinh^2 x = \tfrac{1}{2}(\cosh 2x - 1), \qquad \cosh^2 x = \tfrac{1}{2}(\cosh 2x + 1) \tag{21}$$

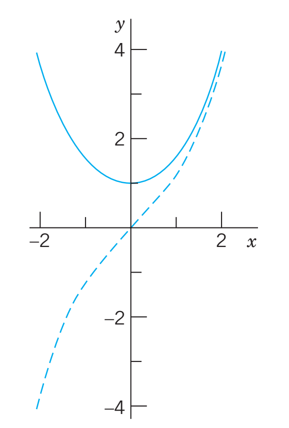


Fig. 551. sinh x (dashed) and cosh x

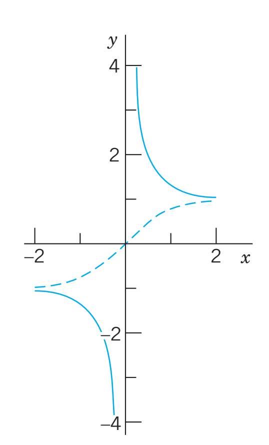


Fig. 552. tanh x (dashed) and coth x

$$
(22) \qquad \begin{cases} \sinh(x \pm y) = \sinh x \cosh y \pm \cosh x \sinh y \\ \cosh(x \pm y) = \cosh x \cosh y \pm \sinh x \sinh y \end{cases}
$$

$$
(23) \qquad \tanh(x \pm y) = \frac{\tanh x \pm \tanh y}{1 \pm \tanh x \tanh y}
$$

Gamma function (Fig. 553 and Table A2 in App. 5). The gamma function $\Gamma(\alpha)$ is defined by the integral

$$
(24) \qquad \Gamma(\alpha) = \int_0^\infty e^{-t} t^{\alpha-1}\, dt \qquad (\alpha > 0),
$$

which is meaningful only if $\alpha > 0$ (or, if we consider complex α, for those α whose real part is positive). Integration by parts gives the important *functional relation of the gamma function,*

$$
(25) \qquad \Gamma(\alpha + 1) = \alpha\Gamma(\alpha).
$$

From (24) we readily have $\Gamma(1) = 1$; hence if α is a positive integer, say k, then by repeated application of (25) we obtain

$$
(26) \qquad \Gamma(k + 1) = k! \qquad (k = 0, 1, \cdots).
$$

This shows that *the gamma function can be regarded as a generalization of the elementary factorial function.* [Sometimes the notation $(\alpha - 1)!$ is used for $\Gamma(\alpha)$, even for noninteger values of α, and the gamma function is also known as the **factorial function**.]

By repeated application of (25) we obtain

$$
\Gamma(\alpha) = \frac{\Gamma(\alpha+1)}{\alpha} = \frac{\Gamma(\alpha+2)}{\alpha(\alpha+1)} = \cdots = \frac{\Gamma(\alpha+k+1)}{\alpha(\alpha+1)(\alpha+2)\cdots(\alpha+k)}
$$

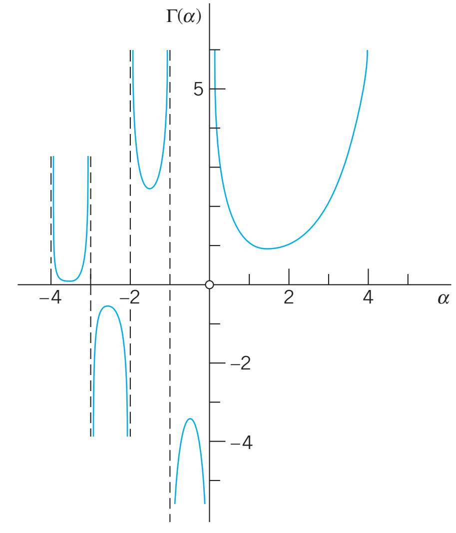


Fig. 553. Gamma function

and we may use this relation

(27) $$\Gamma(\alpha) = \frac{\Gamma(\alpha + k + 1)}{\alpha(\alpha + 1) \cdots (\alpha + k)} \qquad (\alpha \neq 0, -1, -2, \cdots),$$

for defining the gamma function for negative α ($\neq -1, -2, \cdots$), choosing for k the smallest integer such that $\alpha + k + 1 > 0$. *Together with* (24), *this then gives a definition of* $\Gamma(\alpha)$ *for all* α *not equal to zero or a negative integer* (Fig. 553).

It can be shown that the gamma function may also be represented as the limit of a product, namely, by the formula

(28) $$\Gamma(\alpha) = \lim_{n\to\infty} \frac{n!\, n^{\alpha}}{\alpha(\alpha + 1)(\alpha + 2) \cdots (\alpha + n)} \qquad (\alpha \neq 0, -1, \cdots).$$

From (27) or (28) we see that, for complex α, the gamma function $\Gamma(\alpha)$ is a meromorphic function with simple poles at $\alpha = 0, -1, -2, \cdots$.

An approximation of the gamma function for large positive α is given by the **Stirling formula**

(29) $$\Gamma(\alpha + 1) \approx \sqrt{2\pi\alpha}\left(\frac{\alpha}{e}\right)^{\alpha}$$

where e is the base of the natural logarithm. We finally mention the special value

(30) $$\Gamma(\tfrac{1}{2}) = \sqrt{\pi}.$$

Incomplete gamma functions

(31) $$P(\alpha, x) = \int_0^x e^{-t} t^{\alpha - 1}\, dt, \qquad Q(\alpha, x) = \int_x^{\infty} e^{-t} t^{\alpha - 1}\, dt \qquad (\alpha > 0)$$

(32) $$\Gamma(\alpha) = P(\alpha, x) + Q(\alpha, x)$$

Beta function

(33) $$B(x, y) = \int_0^1 t^{x-1}(1 - t)^{y-1}\, dt \qquad (x > 0,\ y > 0)$$

Representation in terms of gamma functions:

(34) $$B(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x + y)}$$

Error function (Fig. 554 and Table A4 in App. 5)

(35) $$\operatorname{erf} x = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

(36) $$\operatorname{erf} x = \frac{2}{\sqrt{\pi}}\left(x - \frac{x^3}{1!3} + \frac{x^5}{2!5} - \frac{x^7}{3!7} + - \cdots\right)$$

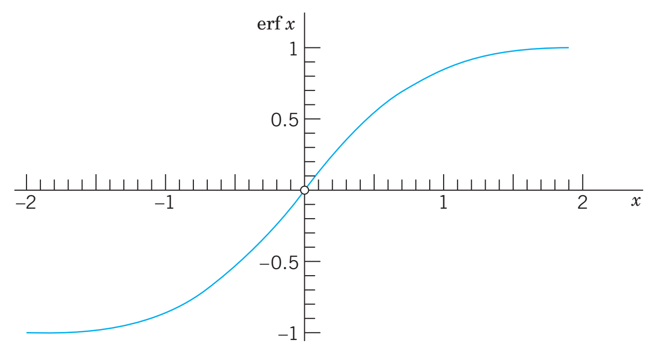


Fig. 554. Error function

erf $(\infty) = 1$, *complementary error function*

$$\text{erfc}\, x = 1 - \text{erf}\, x = \frac{2}{\sqrt{\pi}} \int_x^\infty e^{-t^2}\, dt \tag{37}$$

Fresnel integrals[1] (Fig. 555)

$$C(x) = \int_0^x \cos(t^2)\, dt, \qquad S(x) = \int_0^x \sin(t^2)\, dt \tag{38}$$

$C(\infty) = \sqrt{\pi/8}$, $S(\infty) = \sqrt{\pi/8}$, *complementary functions*

$$\begin{aligned} c(x) &= \sqrt{\frac{\pi}{8}} - C(x) = \int_x^\infty \cos(t^2)\, dt \\ s(x) &= \sqrt{\frac{\pi}{8}} - S(x) = \int_x^\infty \sin(t^2)\, dt \end{aligned} \tag{39}$$

Sine integral (Fig. 556 and Table A4 in App. 5)

$$\text{Si}(x) = \int_0^x \frac{\sin t}{t}\, dt \tag{40}$$

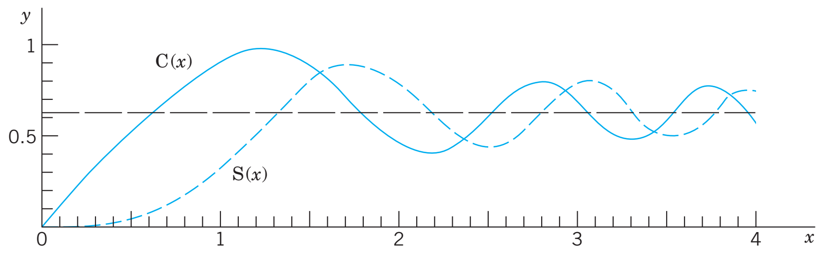


Fig. 555. Fresnel integrals

[1]AUGUSTIN FRESNEL (1788–1827), French physicist and mathematician. For tables see Ref. [GenRef1].

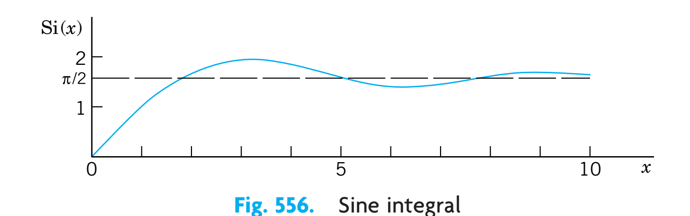


Fig. 556. Sine integral

$\text{Si}(\infty) = \pi/2$, *complementary function*

(41) $$\text{si}(x) = \frac{\pi}{2} - \text{Si}(x) = \int_x^\infty \frac{\sin t}{t}\, dt$$

Cosine integral (Table A4 in App. 5)

(42) $$\text{ci}(x) = \int_x^\infty \frac{\cos t}{t}\, dt \qquad (x > 0)$$

Exponential integral

(43) $$\text{Ei}(x) = \int_x^\infty \frac{e^{-t}}{t}\, dt \qquad (x > 0)$$

Logarithmic integral

(44) $$\text{li}(x) = \int_0^x \frac{dt}{\ln t}$$

A3.2 Partial Derivatives

For differentiation formulas, see inside of front cover.

Let $z = f(x, y)$ be a real function of two independent real variables, x and y. If we keep y constant, say, $y = y_1$, and think of x as a variable, then $f(x, y_1)$ depends on x alone. If the derivative of $f(x, y_1)$ with respect to x for a value $x = x_1$ exists, then the value of this derivative is called the **partial derivative** of $f(x, y)$ *with respect to x at the point* (x_1, y_1) and is denoted by

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1, y_1)} \qquad \text{or by} \qquad \left.\frac{\partial z}{\partial x}\right|_{(x_1, y_1)}.$$

Other notations are

$$f_x(x_1, y_1) \qquad \text{and} \qquad z_x(x_1, y_1);$$

these may be used when subscripts are not used for another purpose and there is no danger of confusion.

We thus have, by the definition of the derivative,

$$\frac{\partial f}{\partial x}\bigg|_{(x_1,y_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1) - f(x_1, y_1)}{\Delta x}. \tag{1}$$

The partial derivative of $z = f(x, y)$ with respect to y is defined similarly; we now keep x constant, say, equal to x_1, and differentiate $f(x_1, y)$ with respect to y. Thus

$$\frac{\partial f}{\partial y}\bigg|_{(x_1,y_1)} = \frac{\partial z}{\partial y}\bigg|_{(x_1,y_1)} = \lim_{\Delta y \to 0} \frac{f(x_1, y_1 + \Delta y) - f(x_1, y_1)}{\Delta y}. \tag{2}$$

Other notations are $f_y(x_1, y_1)$ and $z_y(x_1, y_1)$.

It is clear that the values of those two partial derivatives will in general depend on the point (x_1, y_1). Hence the partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ at a variable point (x, y) are functions of x and y. The function $\partial z/\partial x$ is obtained as in ordinary calculus by differentiating $z = f(x, y)$ with respect to x, ***treating y as a constant***, and $\partial z/\partial y$ is obtained by differentiating z with respect to y, ***treating x as a constant***.

EXAMPLE 1 Let $z = f(x, y) = x^2y + x \sin y$. Then

$$\frac{\partial f}{\partial x} = 2xy + \sin y, \qquad \frac{\partial f}{\partial y} = x^2 + x \cos y.$$ ■

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ of a function $z = f(x, y)$ have a very simple **geometric interpretation**. The function $z = f(x, y)$ can be represented by a surface in space. The equation $y = y_1$ then represents a vertical plane intersecting the surface in a curve, and the partial derivative $\partial z/\partial x$ at a point (x_1, y_1) is the slope of the tangent (that is, $\tan \alpha$ where α is the angle shown in Fig. 557) to the curve. Similarly, the partial derivative $\partial z/\partial y$ at (x_1, y_1) is the slope of the tangent to the curve $x = x_1$ on the surface $z = f(x, y)$ at (x_1, y_1).

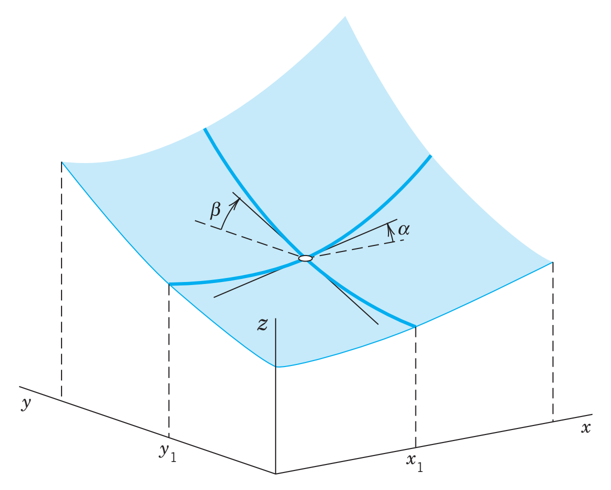


Fig. 557. Geometrical interpretation of first partial derivatives

The partial derivatives $\partial z/\partial x$ and $\partial z/\partial y$ are called *first partial derivatives* or *partial derivatives of first order*. By differentiating these derivatives once more, we obtain the four *second partial derivatives* (or *partial derivatives of second order*)[2]

$$
\begin{aligned}
\frac{\partial^2 f}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) = f_{xx} \\
\frac{\partial^2 f}{\partial x\,\partial y} &= \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) = f_{yx} \\
\frac{\partial^2 f}{\partial y\,\partial x} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) = f_{xy} \\
\frac{\partial^2 f}{\partial y^2} &= \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) = f_{yy}.
\end{aligned} \tag{3}
$$

It can be shown that if all the derivatives concerned are continuous, then the two mixed partial derivatives are equal, so that the order of differentiation does not matter (see Ref. [GenRef4] in App. 1), that is,

$$\frac{\partial^2 z}{\partial x\,\partial y} = \frac{\partial^2 z}{\partial y\,\partial x}. \tag{4}$$

EXAMPLE 2 For the function in Example 1.

$$f_{xx} = 2y, \qquad f_{xy} = 2x + \cos y = f_{yx}, \qquad f_{yy} = -x \sin y. \qquad \blacksquare$$

By differentiating the second partial derivatives again with respect to x and y, respectively, we obtain the *third partial derivatives* or *partial derivatives of the third order* of f, etc.

If we consider a function $f(x, y, z)$ of ***three independent variables***, then we have the three first partial derivatives $f_x(x, y, z)$, $f_y(x, y, z)$, and $f_z(x, y, z)$. Here f_x is obtained by differentiating f with respect to x, ***treating both y and z as constants***. Thus, analogous to (1), we now have

$$\left.\frac{\partial f}{\partial x}\right|_{(x_1, y_1, z_1)} = \lim_{\Delta x \to 0} \frac{f(x_1 + \Delta x, y_1, z_1) - f(x_1, y_1, z_1)}{\Delta x},$$

etc. By differentiating f_x, f_y, f_z again in this fashion we obtain the second partial derivatives of f, etc.

EXAMPLE 3 Let $f(x, y, z) = x^2 + y^2 + z^2 + xy\,e^z$. Then

$$
\begin{aligned}
&f_x = 2x + y\,e^z, && f_y = 2y + x\,e^z, && f_z = 2z + xy\,e^z, \\
&f_{xx} = 2, && f_{xy} = f_{yx} = e^z, && f_{xz} = f_{zx} = y\,e^z, \\
&f_{yy} = 2, && f_{yz} = f_{zy} = x\,e^z, && f_{zz} = 2 + xy\,e^z. \qquad \blacksquare
\end{aligned}
$$

[2] **CAUTION!** In the subscript notation, the subscripts are written in the order in which we differentiate, whereas in the "∂" notation the order is opposite.

A3.3 Sequences and Series

See also Chap. 15.

Monotone Real Sequences

We call a real sequence $x_1, x_2, \cdots, x_n, \cdots$ a **monotone sequence** if it is either **monotone increasing**, that is,

$$x_1 \leqq x_2 \leqq x_3 \leqq \cdots$$

or **monotone decreasing**, that is,

$$x_1 \geqq x_2 \geqq x_3 \geqq \cdots.$$

We call $x_1, x_2, \cdots$ a **bounded sequence** if there is a positive constant K such that $|x_n| < K$ for all n.

THEOREM 1

If a real sequence is bounded and monotone, it converges.

PROOF Let $x_1, x_2, \cdots$ be a bounded monotone increasing sequence. Then its terms are smaller than some number B and, since $x_1 \leqq x_n$ for all n, they lie in the interval $x_1 \leqq x_n \leqq B$, which will be denoted by I_0. We bisect I_0; that is, we subdivide it into two parts of equal length. If the right half (together with its endpoints) contains terms of the sequence, we denote it by I_1. If it does not contain terms of the sequence, then the left half of I_0 (together with its endpoints) is called I_1. This is the first step.

In the second step we bisect I_1, select one half by the same rule, and call it I_2, and so on (see Fig. 558).

In this way we obtain shorter and shorter intervals $I_0, I_1, I_2, \cdots$ with the following properties. Each I_m contains all I_n for $n > m$. No term of the sequence lies to the right of I_m, and, since the sequence is monotone increasing, all x_n with n greater than some number N lie in I_m; of course, N will depend on m, in general. The lengths of the I_m approach zero as m approaches infinity. Hence there is precisely one number, call it L, that lies in all those intervals,[3] and we may now easily prove that the sequence is convergent with the limit L.

In fact, given an $\epsilon > 0$, we choose an m such that the length of I_m is less than ϵ. Then L and all the x_n with $n > N(m)$ lie in I_m, and, therefore, $|x_n - L| < \epsilon$ for all those n. This completes the proof for an increasing sequence. For a decreasing sequence the proof is the same, except for a suitable interchange of "left" and "right" in the construction of those intervals. ■

---

[3]This statement seems to be obvious, but actually it is not; it may be regarded as an axiom of the real number system in the following form. Let $J_1, J_2, \cdots$ be closed intervals such that each J_m contains all J_n with $n > m$, and the lengths of the J_m approach zero as m approaches infinity. Then there is precisely one real number that is contained in all those intervals. This is the so-called **Cantor–Dedekind axiom**, named after the German mathematicians GEORG CANTOR (1845–1918), the creator of set theory, and RICHARD DEDEKIND (1831–1916), known for his fundamental work in number theory. For further details see Ref. [GenRef2] in App. 1. (An interval I is said to be **closed** if its two endpoints are regarded as points belonging to I. It is said to be **open** if the endpoints are not regarded as points of I.)

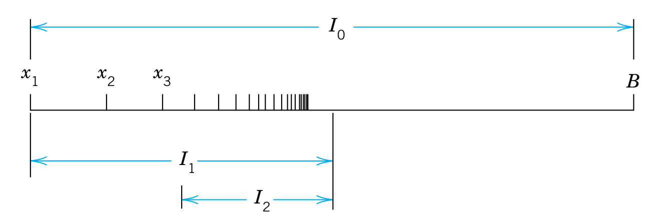


Fig. 558. Proof of Theorem 1

Real Series

THEOREM 2

Leibniz Test for Real Series

Let $x_1, x_2, \cdots$ be real and monotone decreasing to zero, that is,

$$\text{(1)} \qquad \text{(a)} \quad x_1 \geqq x_2 \geqq x_3 \geqq \cdots, \qquad \text{(b)} \quad \lim_{m\to\infty} x_m = 0.$$

Then the series with terms of alternating signs

$$x_1 - x_2 + x_3 - x_4 + - \cdots$$

converges, and for the remainder R_n after the nth term we have the estimate

$$\text{(2)} \qquad |R_n| \leqq x_{n+1}.$$

PROOF Let s_n be the nth partial sum of the series. Then, because of (1a),

$$s_1 = x_1, \qquad s_2 = x_1 - x_2 \leqq s_1,$$

$$s_3 = s_2 + x_3 \geqq s_2, \qquad s_3 = s_1 - (x_2 - x_3) \leqq s_1,$$

so that $s_2 \leqq s_3 \leqq s_1$. Proceeding in this fashion, we conclude that (Fig. 559)

$$\text{(3)} \qquad s_1 \geqq s_3 \geqq s_5 \geqq \cdots \geqq s_6 \geqq s_4 \geqq s_2$$

which shows that the odd partial sums form a bounded monotone sequence, and so do the even partial sums. Hence, by Theorem 1, both sequences converge, say,

$$\lim_{n\to\infty} s_{2n+1} = s, \qquad \lim_{n\to\infty} s_{2n} = s^*.$$

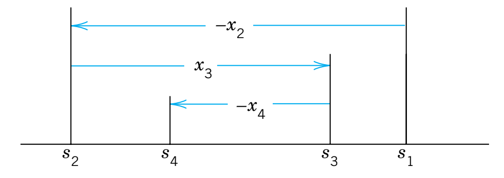


Fig. 559. Proof of the Leibniz test

Now, since $s_{2n+1} - s_{2n} = x_{2n+1}$, we readily see that (lb) implies

$$s - s^* = \lim_{n\to\infty} s_{2n+1} - \lim_{n\to\infty} s_{2n} = \lim_{n\to\infty} (s_{2n+1} - s_{2n}) = \lim_{n\to\infty} x_{2n+1} = 0.$$

Hence $s^* = s$, and the series converges with the sum s.

We prove the estimate (2) for the remainder. Since $s_n \to s$, it follows from (3) that

$$s_{2n+1} \geqq s \geqq s_{2n} \qquad \text{and also} \qquad s_{2n-1} \geqq s \geqq s_{2n}.$$

By subtracting s_{2n} and s_{2n-1}, respectively, we obtain

$$s_{2n+1} - s_{2n} \geqq s - s_{2n} \geqq 0, \qquad 0 \geqq s - s_{2n-1} \geqq s_{2n} - s_{2n-1}.$$

In these inequalities, the first expression is equal to x_{2n+1}, the last is equal to $-x_{2n}$, and the expressions between the inequality signs are the remainders R_{2n} and R_{2n-1}. Thus the inequalities may be written

$$x_{2n+1} \geqq R_{2n} \geqq 0, \qquad 0 \geqq R_{2n-1} \geqq -x_{2n}$$

and we see that they imply (2). This completes the proof. ■

A3.4 Grad, Div, Curl, ∇^2 in Curvilinear Coordinates

To simplify formulas, we write Cartesian coordinates $x = x_1$, $y = x_2$, $z = x_3$. We denote curvilinear coordinates by q_1, q_2, q_3. Through each point P there pass three coordinate surfaces $q_1 = \text{const}$, $q_2 = \text{const}$, $q_3 = \text{const}$. They intersect along coordinate curves. We assume the three coordinate curves through P to be **orthogonal** (perpendicular to each other). We write coordinate transformations as

$$(1) \qquad x_1 = x_1(q_1, q_2, q_3), \qquad x_2 = x_2(q_1, q_2, q_3), \qquad x_3 = x_3(q_1, q_2, q_3).$$

Corresponding transformations of grad, div, curl, and ∇^2 can all be written by using

$$(2) \qquad h_j^2 = \sum_{k=1}^{3} \left(\frac{\partial x_k}{\partial q_j}\right)^2.$$

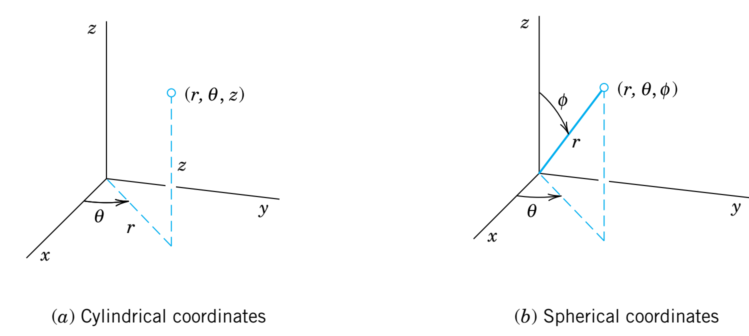
Next to Cartesian coordinates, most important are **cylindrical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = z$ (Fig. 560a) defined by

$$(3) \quad x_1 = q_1 \cos q_2 = r \cos \theta, \qquad x_2 = q_1 \sin q_2 = r \sin \theta, \qquad x_3 = q_3 = z$$

and **spherical coordinates** $q_1 = r$, $q_2 = \theta$, $q_3 = \phi$ (Fig. 560b) defined by[4]

$$(4) \quad \begin{gathered} x_1 = q_1 \cos q_2 \sin q_3 = r \cos\theta \sin\phi, \qquad x_2 = q_1 \sin q_2 \sin q_3 = r \sin\theta \sin\phi \\ x_3 = q_1 \cos q_3 = r \cos\phi. \end{gathered}$$

[4]This is the notation used in calculus and in many other books. It is logical since in it, θ plays the same role as in polar coordinates. **CAUTION!** Some books interchange the roles of θ and ϕ.



(*a*) Cylindrical coordinates (*b*) Spherical coordinates

Fig. 560. Special curvilinear coordinates

In addition to the general formulas for any orthogonal coordinates q_1, q_2, q_3, we shall give additional formulas for these important special cases.

Linear Element *ds*. In Cartesian coordinates,

$$ds^2 = dx_1^2 + dx_2^2 + dx_3^2 \qquad \text{(Sec. 9.5).}$$

For the q-coordinates,

$$ds^2 = h_1^2\, dq_1^2 + h_2^2\, dq_2^2 + h_3^2\, dq_3^2. \tag{5}$$

$$ds^2 = dr^2 + r^2\, d\theta^2 + dz^2 \qquad \text{(Cylindrical coordinates).} \tag{5'}$$

For polar coordinates set $dz^2 = 0$.

$$ds^2 = dr^2 + r^2 \sin^2\phi\, d\theta^2 + r^2\, d\phi^2 \qquad \text{(Spherical coordinates).} \tag{5''}$$

Gradient. grad $f = \nabla f = [f_{x_1},\ f_{x_2},\ f_{x_3}]$ (partial derivatives; Sec. 9.7). In the q-system, with **u, v, w** denoting unit vectors in the positive directions of the q_1, q_2, q_3 coordinate curves, respectively,

$$\operatorname{grad} f = \nabla f = \frac{1}{h_1}\frac{\partial f}{\partial q_1}\mathbf{u} + \frac{1}{h_2}\frac{\partial f}{\partial q_2}\mathbf{v} + \frac{1}{h_3}\frac{\partial f}{\partial q_3}\mathbf{w} \tag{6}$$

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{\partial f}{\partial z}\mathbf{w} \qquad \text{(Cylindrical coordinates)} \tag{6'}$$

$$\operatorname{grad} f = \nabla f = \frac{\partial f}{\partial r}\mathbf{u} + \frac{1}{r\sin\phi}\frac{\partial f}{\partial \theta}\mathbf{v} + \frac{1}{r}\frac{\partial f}{\partial \phi}\mathbf{w} \qquad \text{(Spherical coordinates).} \tag{6''}$$

Divergence div $\mathbf{F} = \nabla \cdot \mathbf{F} = (F_1)_{x_1} + (F_2)_{x_2} + (F_3)_{x_3}$ ($\mathbf{F} = [F_1, F_2, F_3]$, Sec. 9.8);

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{h_1 h_2 h_3}\left[\frac{\partial}{\partial q_1}(h_2 h_3 F_1) + \frac{\partial}{\partial q_2}(h_3 h_1 F_2) + \frac{\partial}{\partial q_3}(h_1 h_2 F_3)\right] \tag{7}$$

$$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{r}\frac{\partial}{\partial r}(rF_1) + \frac{1}{r}\frac{\partial F_2}{\partial \theta} + \frac{\partial F_3}{\partial z} \qquad \text{(Cylindrical coordinates)} \tag{7'}$$

(7″) $$\operatorname{div} \mathbf{F} = \nabla \cdot \mathbf{F} = \frac{1}{r^2} \frac{\partial}{\partial r}(r^2 F_1) + \frac{1}{r \sin \phi} \frac{\partial F_2}{\partial \theta} + \frac{1}{r \sin \phi} \frac{\partial}{\partial \phi}(\sin \phi \, F_3)$$

(Spherical coordinates).

Laplacian $\nabla^2 f = \nabla \cdot \nabla f = \operatorname{div} (\operatorname{grad} f) = f_{x_1 x_1} + f_{x_2 x_2} + f_{x_3 x_3}$ (Sec. 9.8):

(8) $$\nabla^2 f = \frac{1}{h_1 h_2 h_3} \left[\frac{\partial}{\partial q_1} \left(\frac{h_2 h_3}{h_1} \frac{\partial f}{\partial q_1} \right) + \frac{\partial}{\partial q_2} \left(\frac{h_3 h_1}{h_2} \frac{\partial f}{\partial q_2} \right) + \frac{\partial}{\partial q_3} \left(\frac{h_1 h_2}{h_3} \frac{\partial f}{\partial q_3} \right) \right]$$

(8′) $$\nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{1}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \theta^2} + \frac{\partial^2 f}{\partial z^2}$$ (Cylindrical coordinates)

(8″) $$\nabla^2 f = \frac{\partial^2 f}{\partial r^2} + \frac{2}{r} \frac{\partial f}{\partial r} + \frac{1}{r^2 \sin^2 \phi} \frac{\partial^2 f}{\partial \theta^2} + \frac{1}{r^2} \frac{\partial^2 f}{\partial \phi^2} + \frac{\cot \phi}{r^2} \frac{\partial f}{\partial \phi}$$

(Spherical coordinates).

Curl (Sec. 9.9):

(9) $$\operatorname{curl} \mathbf{F} = \nabla \times \mathbf{F} = \frac{1}{h_1 h_2 h_3} \begin{vmatrix} h_1 \mathbf{u} & h_2 \mathbf{v} & h_3 \mathbf{w} \\ \dfrac{\partial}{\partial q_1} & \dfrac{\partial}{\partial q_2} & \dfrac{\partial}{\partial q_3} \\ h_1 F_1 & h_2 F_2 & h_3 F_3 \end{vmatrix} .$$

For cylindrical coordinates we have in (9) (as in the previous formulas)

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 = r, \qquad h_3 = h_z = 1$$

and for spherical coordinates we have

$$h_1 = h_r = 1, \qquad h_2 = h_\theta = q_1 \sin q_3 = r \sin \phi, \qquad h_3 = h_\phi = q_1 = r.$$

APPENDIX 4

Additional Proofs

Section 2.6, page 74

PROOF OF THEOREM 1 **Uniqueness**[1]

Assuming that the problem consisting of the ODE

$$y'' + p(x)y' + q(x)y = 0 \tag{1}$$

and the two initial conditions

$$y(x_0) = K_0, \qquad y'(x_0) = K_1 \tag{2}$$

has two solutions $y_1(x)$ and $y_2(x)$ on the interval I in the theorem, we show that their difference

$$y(x) = y_1(x) - y_2(x)$$

is identically zero on I; then $y_1 \equiv y_2$ on I, which implies uniqueness.

Since (1) is homogeneous and linear, y is a solution of that ODE on I, and since y_1 and y_2 satisfy the same initial conditions, y satisfies the conditions

$$y(x_0) = 0, \qquad y'(x_0) = 0. \tag{11}$$

We consider the function

$$z(x) = y(x)^2 + y'(x)^2$$

and its derivative

$$z' = 2yy' + 2y'y''.$$

From the ODE we have

$$y'' = -py' - qy.$$

By substituting this in the expression for z' we obtain

$$z' = 2yy' - 2py'^2 - 2qyy'. \tag{12}$$

Now, since y and y' are real,

$$(y \pm y')^2 = y^2 \pm 2yy' + y'^2 \geqq 0.$$

[1]This proof was suggested by my colleague, Prof. A. D. Ziebur. In this proof, we use some formula numbers that have not yet been used in Sec. 2.6.

From this and the definition of z we obtain the two inequalities

$$(13) \qquad \text{(a)} \quad 2yy' \leqq y^2 + y'^2 = z, \qquad \text{(b)} \quad -2yy' \leqq y^2 + y'^2 = z.$$

From (13b) we have $2yy' \geqq -z$. Together, $|2yy'| \leqq z$. For the last term in (12) we now obtain

$$-2qyy' \leqq |-2qyy'| = |q||2yy'| \leqq |q|z.$$

Using this result as well as $-p \leqq |p|$ and applying (13a) to the term $2yy'$ in (12), we find

$$z' \leqq z + 2|p|y'^2 + |q|z.$$

Since $y'^2 \leqq y^2 + y'^2 = z$, from this we obtain

$$z' \leqq (1 + 2|p| + |q|)z$$

or, denoting the function in parentheses by h,

$$(14a) \qquad z' \leqq hz \qquad \text{for all } x \text{ on } I.$$

Similarly, from (12) and (13) it follows that

$$(14b) \qquad \begin{aligned} -z' &= -2yy' + 2py'^2 + 2qyy' \\ &\leqq z + 2|p|z + |q|z = hz. \end{aligned}$$

The inequalities (14a) and (14b) are equivalent to the inequalities

$$(15) \qquad z' - hz \leqq 0, \qquad z' + hz \geqq 0.$$

Integrating factors for the two expressions on the left are

$$F_1 = e^{-\int h(x)\,dx} \qquad \text{and} \qquad F_2 = e^{\int h(x)\,dx}.$$

The integrals in the exponents exist because h is continuous. Since F_1 and F_2 are positive, we thus have from (15)

$$F_1(z' - hz) = (F_1 z)' \leqq 0 \qquad \text{and} \qquad F_2(z' + hz) = (F_2 z)' \geqq 0.$$

This means that $F_1 z$ is nonincreasing and $F_2 z$ is nondecreasing on I. Since $z(x_0) = 0$ by (11), when $x \leqq x_0$ we thus obtain

$$F_1 z \geqq (F_1 z)_{x_0} = 0, \qquad F_2 z \leqq (F_2 z)_{x_0} = 0$$

and similarly, when $x \geqq x_0$,

$$F_1 z \leqq 0, \qquad F_2 z \geqq 0.$$

Dividing by F_1 and F_2 and noting that these functions are positive, we altogether have

$$z \leqq 0, \qquad z \geqq 0 \qquad \text{for all } x \text{ on } I.$$

This implies that $z = y^2 + y'^2 \equiv 0$ on I. Hence $y \equiv 0$ or $y_1 \equiv y_2$ on I. ■

Section 5.3, page 182

PROOF OF THEOREM 2 **Frobenius Method. Basis of Solutions. Three Cases**

The formula numbers in this proof are the same as in the text of Sec. 5.3. An additional formula not appearing in Sec. 5.3 will be called (A) (see below).

The ODE in Theorem 2 is

$$y'' + \frac{b(x)}{x}y' + \frac{c(x)}{x^2}y = 0, \tag{1}$$

where $b(x)$ and $c(x)$ are analytic functions. We can write it

$$x^2y'' + xb(x)y' + c(x)y = 0. \tag{1'}$$

The indicial equation of (1) is

$$r(r-1) + b_0 r + c_0 = 0. \tag{4}$$

The roots r_1, r_2 of this quadratic equation determine the general form of a basis of solutions of (1), and there are three possible cases as follows.

Case 1. Distinct Roots Not Differing by an Integer. A first solution of (1) is of the form

$$y_1(x) = x^{r_1}(a_0 + a_1x + a_2x^2 + \cdots) \tag{5}$$

and can be determined as in the power series method. For a proof that in this case, the ODE (1) has a second independent solution of the form

$$y_2(x) = x^{r_2}(A_0 + A_1x + A_2x^2 + \cdots), \tag{6}$$

see Ref. [A11] listed in App. 1.

Case 2. Double Root. The indicial equation (4) has a double root r if and only if $(b_0 - 1)^2 - 4c_0 = 0$, and then $r = \frac{1}{2}(1 - b_0)$. A first solution

$$y_1(x) = x^r(a_0 + a_1x + a_2x^2 + \cdots), \qquad r = \tfrac{1}{2}(1 - b_0), \tag{7}$$

can be determined as in Case 1. We show that a second independent solution is of the form

$$y_2(x) = y_1(x)\ln x + x^r(A_1x + A_2x^2 + \cdots) \qquad (x > 0). \tag{8}$$

We use the method of reduction of order (see Sec. 2.1), that is, we determine $u(x)$ such that $y_2(x) = u(x)y_1(x)$ is a solution of (1). By inserting this and the derivatives

$$y_2' = u'y_1 + uy_1', \qquad y_2'' = u''y_1 + 2u'y_1' + uy_1''$$

into the ODE (1′) we obtain

$$x^2(u''y_1 + 2u'y_1' + uy_1'') + xb(u'y_1 + uy_1') + cuy_1 = 0.$$

Since y_1 is a solution of (1′), the sum of the terms involving u is zero, and this equation reduces to

$$x^2 y_1 u'' + 2x^2 y_1' u' + xby_1 u' = 0.$$

By dividing by $x^2 y_1$ and inserting the power series for b we obtain

$$u'' + \left(2\frac{y_1'}{y_1} + \frac{b_0}{x} + \cdots\right) u' = 0.$$

Here, and in the following, the dots designate terms that are constant or involve positive powers of x. Now, from (7), it follows that

$$\frac{y_1'}{y_1} = \frac{x^{r-1}[ra_0 + (r+1)a_1 x + \cdots]}{x^r[a_0 + a_1 x + \cdots]}$$

$$= \frac{1}{x}\left(\frac{ra_0 + (r+1)a_1 x + \cdots}{a_0 + a_1 x + \cdots}\right) = \frac{r}{x} + \cdots .$$

Hence the previous equation can be written

$$\text{(A)} \qquad u'' + \left(\frac{2r + b_0}{x} + \cdots\right) u' = 0.$$

Since $r = (1 - b_0)/2$, the term $(2r + b_0)/x$ equals $1/x$, and by dividing by u' we thus have

$$\frac{u''}{u'} = -\frac{1}{x} + \cdots .$$

By integration we obtain $\ln u' = -\ln x + \cdots$, hence $u' = (1/x)e^{(\cdots)}$. Expanding the exponential function in powers of x and integrating once more, we see that u is of the form

$$u = \ln x + k_1 x + k_2 x^2 + \cdots .$$

Inserting this into $y_2 = uy_1$, we obtain for y_2 a representation of the form (8).

Case 3. Roots Differing by an Integer. We write $r_1 = r$ and $r_2 = r - p$ where p is a ***positive*** integer. A first solution

$$\text{(9)} \qquad y_1(x) = x^{r_1}(a_0 + a_1 x + a_2 x^2 + \cdots)$$

can be determined as in Cases 1 and 2. We show that a second independent solution is of the form

$$\text{(10)} \qquad y_2(x) = ky_1(x)\ln x + x^{r_2}(A_0 + A_1 x + A_2 x^2 + \cdots)$$

where we may have $k \neq 0$ or $k = 0$. As in Case 2 we set $y_2 = uy_1$. The first steps are literally as in Case 2 and give Eq. (A),

$$u'' + \left(\frac{2r + b_0}{x} + \cdots\right) u' = 0.$$

Now by elementary algebra, the coefficient $b_0 - 1$ of r in (4) equals minus the sum of the roots,

$$b_0 - 1 = -(r_1 + r_2) = -(r + r - p) = -2r + p.$$

Hence $2r + b_0 = p + 1$, and division by u' gives

$$\frac{u''}{u'} = -\left(\frac{p+1}{x} + \cdots\right).$$

The further steps are as in Case 2. Integrating, we find

$$\ln u' = -(p+1)\ln x + \cdots, \qquad \text{thus} \qquad u' = x^{-(p+1)}e^{(\cdots)}$$

where dots stand for some series of nonnegative integer powers of x. By expanding the exponential function as before we obtain a series of the form

$$u' = \frac{1}{x^{p+1}} + \frac{k_1}{x^p} + \cdots + \frac{k_{p-1}}{x^2} + \frac{k_p}{x} + k_{p+1} + k_{p+2}x + \cdots.$$

We integrate once more. Writing the resulting logarithmic term first, we get

$$u = k_p \ln x + \left(-\frac{1}{px^p} - \cdots - \frac{k_{p-1}}{x} + k_{p+1}x + \cdots\right).$$

Hence, by (9) we get for $y_2 = uy_1$ the formula

$$y_2 = k_p y_1 \ln x + x^{r_1 - p}\left(-\frac{1}{p} - \cdots - k_{p-1}x^{p-1} + \cdots\right)(a_0 + a_1 x + \cdots).$$

But this is of the form (10) with $k = k_p$ since $r_1 - p = r_2$ and the product of the two series involves nonnegative integer powers of x only. ■

Section 7.7, page 293

THEOREM

Determinants

The definition of a determinant

$$D = \det \mathbf{A} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{7}$$

as given in Sec. 7.7 is unambiguous, that is, it yields the same value of D no matter which rows or columns we choose in the development.

PROOF In this proof we shall use formula numbers not yet used in Sec. 7.7.

We shall prove first that *the same value is obtained no matter which* **row** *is chosen.*

The proof is by induction. The statement is true for a second-order determinant, for which the developments by the first row $a_{11}a_{22} + a_{12}(-a_{21})$ and by the second row $a_{21}(-a_{12}) + a_{22}a_{11}$ give the same value $a_{11}a_{22} - a_{12}a_{21}$. Assuming the statement to be true for an $(n - 1)$st-order determinant, we prove that it is true for an nth-order determinant.

For this purpose we expand D in terms of each of two arbitrary rows, say, the ith and the jth, and compare the results. Without loss of generality let us assume $i < j$.

First Expansion. We expand D by the ith row. A typical term in this expansion is

$$a_{ik}C_{ik} = a_{ik} \cdot (-1)^{i+k} M_{ik}. \tag{19}$$

The minor M_{ik} of a_{ik} in D is an $(n - 1)$st-order determinant. By the induction hypothesis we may expand it by any row. We expand it by the row corresponding to the jth row of D. This row contains the entries a_{jl} $(l \neq k)$. It is the $(j - 1)$st row of M_{ik}, because M_{ik} does not contain entries of the ith row of D, and $i < j$. We have to distinguish between two cases as follows.

Case I. If $l < k$, then the entry a_{jl} belongs to the lth column of M_{ik} (see Fig. 561). Hence the term involving a_{jl} in this expansion is

$$a_{jl} \cdot (\text{cofactor of } a_{jl} \text{ in } M_{ik}) = a_{jl} \cdot (-1)^{(j-1)+l} M_{ikjl} \tag{20}$$

where M_{ikjl} is the minor of a_{jl} in M_{ik}. Since this minor is obtained from M_{ik} by deleting the row and column of a_{jl}, it is obtained from D by deleting the ith and jth rows and the kth and lth columns of D. We insert the expansions of the M_{ik} into that of D. Then it follows from (19) and (20) that the terms of the resulting representation of D are of the form

$$a_{ik}a_{jl} \cdot (-1)^b M_{ikjl} \qquad (l < k) \tag{21a}$$

where

$$b = i + k + j + l - 1.$$

Case II. If $l > k$, the only difference is that then a_{jl} belongs to the $(l - 1)$st column of M_{ik}, because M_{ik} does not contain entries of the kth column of D, and $k < l$. This causes an additional minus sign in (20), and, instead of (21a), we therefore obtain

$$-a_{ik}a_{jl} \cdot (-1)^b M_{ikjl} \qquad (l > k) \tag{21b}$$

where b is the same as before.

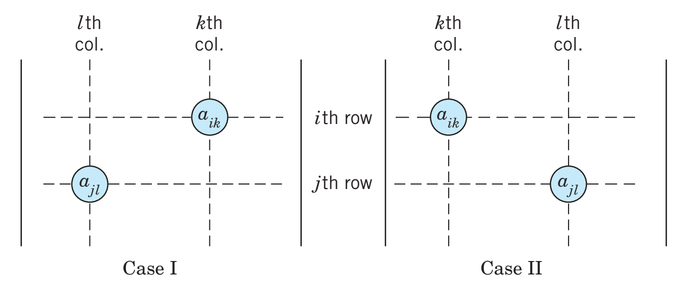


Fig. 561. Cases I and II of the two expansions of D

Second Expansion. We now expand D at first by the jth row. A typical term in this expansion is

$$a_{jl}C_{jl} = a_{jl} \cdot (-1)^{j+l}M_{jl}. \tag{22}$$

By the induction hypothesis we may expand the minor M_{jl} of a_{jl} in D by its ith row, which corresponds to the ith row of D, since $j > i$.

Case I. If $k > l$, the entry a_{ik} in that row belongs to the $(k-1)$st column of M_{jl}, because M_{jl} does not contain entries of the lth column of D, and $l < k$ (see Fig. 561). Hence the term involving a_{ik} in this expansion is

$$a_{ik} \cdot (\text{cofactor of } a_{ik} \text{ in } M_{jl}) = a_{ik} \cdot (-1)^{i+(k-1)}M_{ikjl}, \tag{23}$$

where the minor M_{ikjl} of a_{ik} in M_{jl} is obtained by deleting the ith and jth rows and the kth and lth columns of D [and is, therefore, identical with M_{ikjl} in (20), so that our notation is consistent]. We insert the expansions of the M_{jl} into that of D. It follows from (22) and (23) that this yields a representation whose terms are identical with those given by (21a) when $l < k$.

Case II. If $k < l$, then a_{ik} belongs to the kth column of M_{jl}, we obtain an additional minus sign, and the result agrees with that characterized by (21b).

We have shown that the two expansions of D consist of the same terms, and this proves our statement concerning rows.

The proof of the statement concerning ***columns*** is quite similar; if we expand D in terms of two arbitrary columns, say, the kth and the lth, we find that the general term involving $a_{jl}a_{ik}$ is exactly the same as before. This proves that not only all column expansions of D yield the same value, but also that their common value is equal to the common value of the row expansions of D.

This completes the proof and shows that *our definition of an nth-order determinant is unambiguous.* ■

Section 9.3, page 368

PROOF OF FORMULA (2)

We prove that in right-handed Cartesian coordinates, the vector product

$$\mathbf{v} = \mathbf{a} \times \mathbf{b} = [a_1, \quad a_2, \quad a_3] \times [b_1, \quad b_2, \quad b_3]$$

has the components

$$v_1 = a_2b_3 - a_3b_2, \qquad v_2 = a_3b_1 - a_1b_3, \qquad v_3 = a_1b_2 - a_2b_1. \tag{2}$$

We need only consider the case $\mathbf{v} \neq \mathbf{0}$. Since $\mathbf{v}$ is perpendicular to both $\mathbf{a}$ and $\mathbf{b}$, Theorem 1 in Sec. 9.2 gives $\mathbf{a} \cdot \mathbf{v} = 0$ and $\mathbf{b} \cdot \mathbf{v} = 0$; in components [see (2), Sec. 9.2],

$$\begin{aligned} a_1v_1 + a_2v_2 + a_3v_3 &= 0 \\ b_1v_1 + b_2v_2 + b_3v_3 &= 0. \end{aligned} \tag{3}$$

Multiplying the first equation by b_3, the last by a_3, and subtracting, we obtain

$$(a_3b_1 - a_1b_3)v_1 = (a_2b_3 - a_3b_2)v_2.$$

Multiplying the first equation by b_1, the last by a_1, and subtracting, we obtain

$$(a_1b_2 - a_2b_1)v_2 = (a_3b_1 - a_1b_3)v_3.$$

We can easily verify that these two equations are satisfied by

$$\text{(4)} \qquad v_1 = c(a_2b_3 - a_3b_2), \qquad v_2 = c(a_3b_1 - a_1b_3), \qquad v_3 = c(a_1b_2 - a_2b_1)$$

where c is a constant. The reader may verify, by inserting, that (4) also satisfies (3). Now each of the equations in (3) represents a plane through the origin in $v_1v_2v_3$-space. The vectors **a** and **b** are normal vectors of these planes (see Example 6 in Sec. 9.2). Since $\mathbf{v} \neq \mathbf{0}$, these vectors are not parallel and the two planes do not coincide. Hence their intersection is a straight line L through the origin. Since (4) is a solution of (3) and, for varying c, represents a straight line, we conclude that (4) represents L, and every solution of (3) must be of the form (4). In particular, the components of **v** must be of this form, where c is to be determined. From (4) we obtain

$$|\mathbf{v}|^2 = v_1^2 + v_2^2 + v_3^2 = c^2\left[(a_2b_3 - a_3b_2)^2 + (a_3b_1 - a_1b_3)^2 + (a_1b_2 - a_2b_1)^2\right].$$

This can be written

$$|\mathbf{v}|^2 = c^2\left[(a_1^2 + a_2^2 + a_3^2)(b_1^2 + b_2^2 + b_3^2) - (a_1b_1 + a_2b_2 + a_3b_3)^2\right],$$

as can be verified by performing the indicated multiplications in both formulas and comparing. Using (2) in Sec. 9.2, we thus have

$$|\mathbf{v}|^2 = c^2\left[(\mathbf{a} \cdot \mathbf{a})(\mathbf{b} \cdot \mathbf{b}) - (\mathbf{a} \cdot \mathbf{b})^2\right].$$

By comparing this with formula (12) in Prob. 4 of Problem Set 9.3 we conclude that $c = \pm 1$.

We show that $c = +1$. This can be done as follows.

If we change the lengths and directions of **a** and **b** continuously and so that at the end $\mathbf{a} = \mathbf{i}$ and $\mathbf{b} = \mathbf{j}$ (Fig. 188a in Sec. 9.3), then **v** will change its length and direction continuously, and at the end, $\mathbf{v} = \mathbf{i} \times \mathbf{j} = \mathbf{k}$. Obviously we may effect the change so that both **a** and **b** remain different from the zero vector and are not parallel at any instant. Then **v** is never equal to the zero vector, and since the change is continuous and c can only assume the values $+1$ or -1, it follows that at the end c must have the same value as before. Now at the end $\mathbf{a} = \mathbf{i}$, $\mathbf{b} = \mathbf{j}$, $\mathbf{v} = \mathbf{k}$ and, therefore, $a_1 = 1$, $b_2 = 1$, $v_3 = 1$, and the other components in (4) are zero. Hence from (4) we see that $v_3 = c = +1$. This proves Theorem 1.

For a left-handed coordinate system, $\mathbf{i} \times \mathbf{j} = -\mathbf{k}$ (see Fig. 188b in Sec. 9.3), resulting in $c = -1$. This proves the statement right after formula (2). ■

Section 9.9, page 408

PROOF OF THE INVARIANCE OF THE CURL

This proof will follow from two theorems (A and B), which we prove first.

THEOREM A

Transformation Law for Vector Components

For any vector **v** *the components* v_1, v_2, v_3 *and* v_1^*, v_2^*, v_3^* *in any two systems of Cartesian coordinates* x_1, x_2, x_3 *and* x_1^*, x_2^*, x_3^*, *respectively, are related by*

$$\begin{aligned} v_1^* &= c_{11}v_1 + c_{12}v_2 + c_{13}v_3 \\ v_2^* &= c_{21}v_1 + c_{22}v_2 + c_{23}v_3 \\ v_3^* &= c_{31}v_1 + c_{32}v_2 + c_{33}v_3, \end{aligned} \tag{1}$$

and conversely

$$\begin{aligned} v_1 &= c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^* \\ v_2 &= c_{12}v_1^* + c_{22}v_2^* + c_{32}v_3^* \\ v_3 &= c_{13}v_1^* + c_{23}v_2^* + c_{33}v_3^* \end{aligned} \tag{2}$$

with coefficients

$$\begin{array}{lll} c_{11} = \mathbf{i}^* \bullet \mathbf{i} & c_{12} = \mathbf{i}^* \bullet \mathbf{j} & c_{13} = \mathbf{i}^* \bullet \mathbf{k} \\ c_{21} = \mathbf{j}^* \bullet \mathbf{i} & c_{22} = \mathbf{j}^* \bullet \mathbf{j} & c_{23} = \mathbf{j}^* \bullet \mathbf{k} \\ c_{31} = \mathbf{k}^* \bullet \mathbf{i} & c_{32} = \mathbf{k}^* \bullet \mathbf{j} & c_{33} = \mathbf{k}^* \bullet \mathbf{k} \end{array} \tag{3}$$

satisfying

$$\sum_{j=1}^{3} c_{kj}c_{mj} = \delta_{km} \qquad (k, m = 1, 2, 3), \tag{4}$$

where the **Kronecker delta**[2] *is given by*

$$\delta_{km} = \begin{cases} 0 & (k \neq m) \\ 1 & (k = m) \end{cases}$$

and **i, j, k** *and* **i*, j*, k*** *denote the unit vectors in the positive* x_1-, x_2-, x_3- *and* x_1^*-, x_2^*-, x_3^*-*directions, respectively.*

---

[2]LEOPOLD KRONECKER (1823–1891), German mathematician at Berlin, who made important contributions to algebra, group theory, and number theory.

We shall keep our discussion completely independent of Chap. 7, but readers familiar with matrices should recognize that we are dealing with **orthogonal transformations and matrices** and that our present theorem follows from Theorem 2 in Sec. 8.3.

PROOF The representation of $\mathbf{v}$ in the two systems are

$$\text{(5)} \qquad \text{(a)} \quad \mathbf{v} = v_1\mathbf{i} + v_2\mathbf{j} + v_3\mathbf{k} \qquad \text{(b)} \quad \mathbf{v} = v_1^*\mathbf{i}^* + v_2^*\mathbf{j}^* + v_3^*\mathbf{k}^*.$$

Since $\mathbf{i}^* \cdot \mathbf{i}^* = 1$, $\mathbf{i}^* \cdot \mathbf{j}^* = 0$, $\mathbf{i}^* \cdot \mathbf{k}^* = 0$, we get from (5b) simply $\mathbf{i}^* \cdot \mathbf{v} = v_1^*$ and from this and (5a)

$$v_1^* = \mathbf{i}^* \cdot \mathbf{v} = \mathbf{i}^* \cdot v_1\mathbf{i} + \mathbf{i}^* \cdot v_2\mathbf{j} + \mathbf{i}^* \cdot v_3\mathbf{k} = v_1\mathbf{i}^* \cdot \mathbf{i} + v_2\mathbf{i}^* \cdot \mathbf{j} + v_3\mathbf{i}^* \cdot \mathbf{k}.$$

Because of (3), this is the first formula in (1), and the other two formulas are obtained similarly, by considering $\mathbf{j}^* \cdot \mathbf{v}$ and then $\mathbf{k}^* \cdot \mathbf{v}$. Formula (2) follows by the same idea, taking $\mathbf{i} \cdot \mathbf{v} = v_1$ from (5a) and then from (5b) and (3)

$$v_1 = \mathbf{i} \cdot \mathbf{v} = v_1^*\mathbf{i} \cdot \mathbf{i}^* + v_2^*\mathbf{i} \cdot \mathbf{j}^* + v_3^*\mathbf{i} \cdot \mathbf{k}^* = c_{11}v_1^* + c_{21}v_2^* + c_{31}v_3^*,$$

and similarly for the other two components.

We prove (4). We can write (1) and (2) briefly as

$$\text{(6)} \qquad \text{(a)} \quad v_j = \sum_{m=1}^{3} c_{mj}v_m^*, \qquad \text{(b)} \quad v_k^* = \sum_{j=1}^{3} c_{kj}v_j.$$

Substituting v_j into v_k^*, we get

$$v_k^* = \sum_{j=1}^{3} c_{kj} \sum_{m=1}^{3} c_{mj}v_m^* = \sum_{m=1}^{3} v_m^* \left(\sum_{j=1}^{3} c_{kj}c_{mj} \right),$$

where $k = 1, 2, 3$. Taking $k = 1$, we have

$$v_1^* = v_1^* \left(\sum_{j=1}^{3} c_{1j}c_{1j} \right) + v_2^* \left(\sum_{j=1}^{3} c_{1j}c_{2j} \right) + v_3^* \left(\sum_{j=1}^{3} c_{1j}c_{3j} \right).$$

For this to hold for *every* vector $\mathbf{v}$, the first sum must be 1 and the other two sums 0. This proves (4) with $k = 1$ for $m = 1, 2, 3$. Taking $k = 2$ and then $k = 3$, we obtain (4) with $k = 2$ and 3, for $m = 1, 2, 3$. ■

THEOREM B

Transformation Law for Cartesian Coordinates

The transformation of any Cartesian $x_1x_2x_3$-coordinate system into any other Cartesian $x_1^*x_2^*x_3^*$-coordinate system is of the form

$$\text{(7)} \qquad x_m^* = \sum_{j=1}^{3} c_{mj}x_j + b_m, \quad m = 1, 2, 3,$$

with coefficients (3) *and constants b_1, b_2, b_3; conversely,*

$$\text{(8)} \qquad x_k = \sum_{n=1}^{3} c_{nk}x_n^* + \tilde{b}_k, \qquad k = 1, 2, 3.$$

Theorem B follows from Theorem A by noting that the most general transformation of a Cartesian coordinate system into another such system may be decomposed into a transformation of the type just considered and a translation; and under a translation, corresponding coordinates differ merely by a constant.

PROOF OF THE INVARIANCE OF THE CURL

We write again x_1, x_2, x_3 instead of x, y, z, and similarly x_1^*, x_2^*, x_3^* for other Cartesian coordinates, assuming that both systems are right-handed. Let a_1, a_2, a_3 denote the components of curl $\mathbf{v}$ in the $x_1x_2x_3$-coordinates, as given by (1), Sec. 9.9, with

$$x = x_1, \qquad y = x_2, \qquad z = x_3.$$

Similarly, let a_1^*, a_2^*, a_3^* denote the components of curl $\mathbf{v}$ in the $x_1^*x_2^*x_3^*$-coordinate system. We prove that the length and direction of curl $\mathbf{v}$ are independent of the particular choice of Cartesian coordinates, as asserted. We do this by showing that the components of curl $\mathbf{v}$ satisfy the transformation law (2), which is characteristic of vector components. We consider a_1. We use (6a), and then the chain rule for functions of several variables (Sec. 9.6). This gives

$$a_1 = \frac{\partial v_3}{\partial x_2} - \frac{\partial v_2}{\partial x_3} = \sum_{m=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_3} \right)$$

$$= \sum_{m=1}^{3} \sum_{j=1}^{3} \left(c_{m3} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_2} - c_{m2} \frac{\partial v_m^*}{\partial x_j^*} \frac{\partial x_j^*}{\partial x_3} \right).$$

From this and (7) we obtain

$$a_1 = \sum_{m=1}^{3} \sum_{j=1}^{3} (c_{m3}c_{j2} - c_{m2}c_{j3}) \frac{\partial v_m^*}{\partial x_j^*}$$

$$= (c_{33}c_{22} - c_{32}c_{23}) \left(\frac{\partial v_3^*}{\partial x_2^*} - \frac{\partial v_2^*}{\partial x_3^*} \right) + \cdots$$

$$= (c_{33}c_{22} - c_{32}c_{23})a_1^* + (c_{13}c_{32} - c_{12}c_{33})a_2^* + (c_{23}c_{12} - c_{22}c_{13})a_3^*.$$

Note what we did. The double sum had $3 \times 3 = 9$ terms, 3 of which were zero (when $m = j$), and the remaining 6 terms we combined in pairs as we needed them in getting a_1^*, a_2^*, a_3^*.

We now use (3), Lagrange's identity (see Formula (15) in Team Project 24 in Problem Set 9.3) and $\mathbf{k}^* \times \mathbf{j}^* = -\mathbf{i}^*$ and $\mathbf{k} \times \mathbf{j} = -\mathbf{i}$. Then

$$c_{33}c_{22} - c_{32}c_{23} = (\mathbf{k}^* \cdot \mathbf{k})(\mathbf{j}^* \cdot \mathbf{j}) - (\mathbf{k}^* \cdot \mathbf{j})(\mathbf{j}^* \cdot \mathbf{k})$$

$$= (\mathbf{k}^* \times \mathbf{j}^*) \cdot (\mathbf{k} \times \mathbf{j}) = \mathbf{i}^* \cdot \mathbf{i} = c_{11}, \qquad \text{etc.}$$

Hence $a_1 = c_{11}a_1^* + c_{21}a_2^* + c_{31}a_3^*$. This is of the form of the first formula in (2) in Theorem A, and the other two formulas of the form (2) are obtained similarly. This proves the theorem for right-handed systems. If the $x_1x_2x_3$-coordinates are left-handed, then $\mathbf{k} \times \mathbf{j} = +\mathbf{i}$, but then there is a minus sign in front of the determinant in (1), Sec. 9.9. ■

Section 10.2, page 420

PROOF OF THEOREM 1, PART (b) We prove that if

$$\int_C \mathbf{F}(\mathbf{r}) \cdot d\mathbf{r} = \int_C (F_1\,dx + F_2\,dy + F_3\,dz) \tag{1}$$

with continuous F_1, F_2, F_3 in a domain D is independent of path in D, then $F = \text{grad } f$ in D for some f; in components

$$F_1 = \frac{\partial f}{\partial x}, \qquad F_2 = \frac{\partial f}{\partial y}, \qquad F_3 = \frac{\partial f}{\partial z}. \tag{2'}$$

We choose any fixed A: (x_0, y_0, z_0) in D and any B: (x, y, z) in D and define f by

$$f(x, y, z) = f_0 + \int_A^B (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*) \tag{3}$$

with any constant f_0 and any path from A to B in D. Since A is fixed and we have independence of path, the integral depends only on the coordinates x, y, z, so that (3) defines a function $f(x, y, z)$ in D. We show that $\mathbf{F} = \text{grad } f$ with this f, beginning with the first of the three relations (2′). Because of independence of path we may integrate from A to B_1: (x_1, y, z) and then parallel to the x-axis along the segment B_1B in Fig. 562 with B_1 chosen so that the whole segment lies in D. Then

$$f(x, y, z) = f_0 + \int_A^{B_1} (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*) + \int_{B_1}^{B} (F_1\,dx^* + F_2\,dy^* + F_3\,dz^*).$$

We now take the partial derivative with respect to x on both sides. On the left we get $\partial f/\partial x$. We show that on the right we get F_1. The derivative of the first integral is zero because A: (x_0, y_0, z_0) and B_1: (x_1, y, z) do not depend on x. We consider the second integral. Since on the segment B_1B, both y and z are constant, the terms $F_2\,dy^*$ and

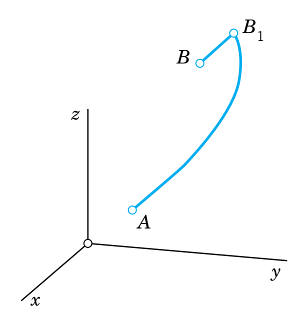


Fig. 562. Proof of Theorem 1

$F_3\, dz^*$ do not contribute to the derivative of the integral. The remaining part can be written as a definite integral,

$$\int_{B_1}^{B} F_1\, dx^* = \int_{x_1}^{x} F_1(x^*, y, z)\, dx^*.$$

Hence its partial derivative with respect to x is $F_1(x, y, z)$, and the first of the relations $(2')$ is proved. The other two formulas in $(2')$ follow by the same argument. ■

Section 11.5, page 500

THEOREM

Reality of Eigenvalues

If p, q, r, and p' in the Sturm–Liouville equation (1) *of Sec.* 11.5 *are real-valued and continuous on the interval $a \leqq x \leqq b$ and $r(x) > 0$ throughout that interval (or $r(x) < 0$ throughout that interval), then all the eigenvalues of the Sturm–Liouville problem* (1), (2), *Sec.* 11.5, *are real.*

PROOF Let $\lambda = \alpha + i\beta$ be an eigenvalue of the problem and let

$$y(x) = u(x) + iv(x)$$

be a corresponding eigenfunction; here α, β, u, and v are real. Substituting this into (1), Sec. 11.5, we have

$$(pu' + ipv')' + (q + \alpha r + i\beta r)(u + iv) = 0.$$

This complex equation is equivalent to the following pair of equations for the real and the imaginary parts:

$$(pu')' + (q + \alpha r)u - \beta rv = 0$$
$$(pv')' + (q + \alpha r)v + \beta ru = 0.$$

Multiplying the first equation by v, the second by $-u$ and adding, we get

$$\begin{aligned} -\beta(u^2 + v^2)r &= u(pv')' - v(pu')' \\ &= [(pv')u - (pu')v]'. \end{aligned}$$

The expression in brackets is continuous on $a \leqq x \leqq b$, for reasons similar to those in the proof of Theorem 1, Sec. 11.5. Integrating over x from a to b, we thus obtain

$$-\beta \int_a^b (u^2 + v^2)r\, dx = \Big[p(uv' - u'v) \Big]_a^b.$$

Because of the boundary conditions, the right side is zero; this is as in that proof. Since y is an eigenfunction, $u^2 + v^2 \not\equiv 0$. Since y and r are continuous and $r > 0$ (or $r < 0$) on the interval $a \leqq x \leqq b$, the integral on the left is not zero. Hence, $\beta = 0$, which means that $\lambda = \alpha$ is real. This completes the proof. ■

Section 13.4, page 627

PROOF OF THEOREM 2 Cauchy–Riemann Equations

We prove that Cauchy–Riemann equations

$$u_x = v_y, \qquad u_y = -v_x \tag{1}$$

are sufficient for a complex function $f(z) = u(x, y) + iv(x, y)$ to be analytic; precisely, *if the real part u and the imaginary part v of f(z) satisfy* (1) *in a domain D in the complex plane and if the partial derivatives in* (1) *are* ***continuous*** *in D, then f(z) is analytic in D.*

In this proof we write $\Delta z = \Delta x + i\Delta y$ and $\Delta f = f(z + \Delta z) - f(z)$. The idea of proof is as follows.

(a) We express Δf in terms of first partial derivatives of u and v, by applying the mean value theorem of Sec. 9.6.

(b) We get rid of partial derivatives with respect to y by applying the Cauchy–Riemann equations.

(c) We let Δz approach zero and show that then $\Delta f/\Delta z$, as obtained, approaches a limit, which is equal to $u_x + iv_x$, the right side of (4) in Sec. 13.4, regardless of the way of approach to zero.

(a) Let P: (x, y) be any fixed point in D. Since D is a domain, it contains a neighborhood of P. We can choose a point Q: $(x + \Delta x, y + \Delta y)$ in this neighborhood such that the straight-line segment PQ is in D. Because of our continuity assumptions we may apply the mean value theorem in Sec. 9.6. This yields

$$u(x + \Delta x, y + \Delta y) - u(x, y) = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1)$$
$$v(x + \Delta x, y + \Delta y) - v(x, y) = (\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)$$

where M_1 and M_2 ($\neq M_1$ in general!) are suitable points on that segment. The first line is Re Δf and the second is Im Δf, so that

$$\Delta f = (\Delta x)u_x(M_1) + (\Delta y)u_y(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)v_y(M_2)].$$

(b) $u_y = -v_x$ and $v_y = u_x$ by the Cauchy–Riemann equations, so that

$$\Delta f = (\Delta x)u_x(M_1) - (\Delta y)v_x(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)].$$

Also $\Delta z = \Delta x + i\Delta y$, so that we can write $\Delta x = \Delta z - i\Delta y$ in the first term and $\Delta y = (\Delta z - \Delta x)/i = -i(\Delta z - \Delta x)$ in the second term. This gives

$$\Delta f = (\Delta z - i\Delta y)u_x(M_1) + i(\Delta z - \Delta x)v_x(M_1) + i[(\Delta x)v_x(M_2) + (\Delta y)u_x(M_2)].$$

By performing the multiplications and reordering we obtain

$$\Delta f = (\Delta z)u_x(M_1) - i\Delta y\{u_x(M_1) - u_x(M_2)\}$$
$$+ i[(\Delta z)v_x(M_1) - \Delta x\{v_x(M_1) - v_x(M_2)\}].$$

Division by Δz now yields

$$\text{(A)} \quad \frac{\Delta f}{\Delta z} = u_x(M_1) + iv_x(M_1) - \frac{i\Delta y}{\Delta z}\{u_x(M_1) - u_x(M_2)\} - \frac{i\Delta x}{\Delta z}\{v_x(M_1) - v_x(M_2)\}.$$

(c) We finally let Δz approach zero and note that $|\Delta y/\Delta z| \leqq 1$ and $|\Delta x/\Delta z| \leqq 1$ in (A). Then Q: $(x + \Delta x, y + \Delta y)$ approaches P: (x, y), so that M_1 and M_2 must approach P. Also, since the partial derivatives in (A) are assumed to be continuous, they approach their value at P. In particular, the differences in the braces $\{\cdot \cdot \cdot\}$ in (A) approach zero. Hence the limit of the right side of (A) exists and is independent of the path along which $\Delta z \to 0$. We see that this limit equals the right side of (4) in Sec. 13.4. This means that $f(z)$ is analytic at every point z in D, and the proof is complete. ■

Section 14.2, pages 653–654

GOURSAT'S PROOF OF CAUCHY'S INTEGRAL THEOREM Goursat proved Cauchy's integral theorem without assuming that $f'(z)$ is continuous, as follows.

We start with the case when C is the boundary of a triangle. We orient C counterclockwise. By joining the midpoints of the sides we subdivide the triangle into four congruent triangles (Fig. 563). Let $C_{\text{I}}, C_{\text{II}}, C_{\text{III}}, C_{\text{IV}}$ denote their boundaries. We claim that (see Fig. 563).

$$\text{(1)} \qquad \oint_C f\,dz = \oint_{C_{\text{I}}} f\,dz + \oint_{C_{\text{II}}} f\,dz + \oint_{C_{\text{III}}} f\,dz + \oint_{C_{\text{IV}}} f\,dz.$$

Indeed, on the right we integrate along each of the three segments of subdivision in both possible directions (Fig. 563), so that the corresponding integrals cancel out in pairs, and the sum of the integrals on the right equals the integral on the left. We now pick an integral on the right that is biggest in absolute value and call its path C_1. Then, by the triangle inequality (Sec. 13.2),

$$\left|\oint_C f\,dz\right| \leqq \left|\oint_{C_{\text{I}}} f\,dz\right| + \left|\oint_{C_{\text{II}}} f\,dz\right| + \left|\oint_{C_{\text{III}}} f\,dz\right| + \left|\oint_{C_{\text{IV}}} f\,dz\right| \leqq 4\left|\oint_{C_1} f\,dz\right|.$$

We now subdivide the triangle bounded by C_1 as before and select a triangle of subdivision with boundary C_2 for which

$$\left|\oint_{C_1} f\,dz\right| \leqq 4\left|\oint_{C_2} f\,dz\right|. \qquad \text{Then} \qquad \left|\oint_C f\,dz\right| \leqq 4^2\left|\oint_{C_2} f\,dz\right|.$$

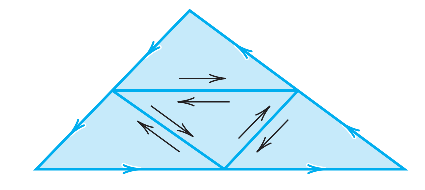


Fig. 563. Proof of Cauchy's integral theorem

Continuing in this fashion, we obtain a sequence of triangles $T_1, T_2, \cdots$ with boundaries $C_1, C_2, \cdots$ that are similar and such that T_n lies in T_m when $n > m$, and

$$\left|\oint_C f\,dz\right| \leqq 4^n \left|\oint_{C_n} f\,dz\right|, \qquad n = 1, 2, \cdots. \tag{2}$$

Let z_0 be the point that belongs to all these triangles. Since f is differentiable at $z = z_0$, the derivative $f'(z_0)$ exists. Let

$$h(z) = \frac{f(z) - f(z_0)}{z - z_0} - f'(z_0). \tag{3}$$

Solving this algebraically for $f(z)$ we have

$$f(z) = f(z_0) + (z - z_0)f'(z_0) + h(z)(z - z_0).$$

Integrating this over the boundary C_n of the triangle T_n gives

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} f(z_0)\,dz + \oint_{C_n} (z - z_0)f'(z_0)\,dz + \oint_{C_n} h(z)(z - z_0)dz.$$

Since $f(z_0)$ and $f'(z_0)$ are constants and C_n is a closed path, the first two integrals on the right are zero, as follows from Cauchy's proof, which is applicable because the integrands do have continuous derivatives (0 and const, respectively). We thus have

$$\oint_{C_n} f(z)\,dz = \oint_{C_n} h(z)(z - z_0)\,dz.$$

Since $f'(z_0)$ is the limit of the difference quotient in (3), for given $\epsilon > 0$ we can find a $\delta > 0$ such that

$$|h(z)| < \epsilon \qquad \text{when} \qquad |z - z_0| < \delta. \tag{4}$$

We may now take n so large that the triangle T_n lies in the disk $|z - z_0| < \delta$. Let L_n be the length of C_n. Then $|z - z_0| < L_n$ for all z on C_n and z_0 in T_n. From this and (4) we have $|h(z)(z - z_0)| < \epsilon L_n$. The *ML*-inequality in Sec. 14.1 now gives

$$\left|\oint_{C_n} f(z)\,dz\right| = \left|\oint_{C_n} h(z)(z - z_0)\,dz\right| \leqq \epsilon L_n \cdot L_n = \epsilon L_n^2. \tag{5}$$

Now denote the length of C by L. Then the path C_1 has the length $L_1 = L/2$, the path C_2 has the length $L_2 = L_1/2 = L/4$, etc., and C_n has the length $L_n = L/2^n$. Hence $L_n^2 = L^2/4^n$. From (2) and (5) we thus obtain

$$\left|\oint_C f\,dz\right| \leqq 4^n \left|\oint_{C_n} f\,dz\right| \leqq 4^n \epsilon L_n^2 = 4^n \epsilon \frac{L^2}{4^n} = \epsilon L^2.$$

By choosing ϵ (> 0) sufficiently small we can make the expression on the right as small as we please, while the expression on the left is the definite value of an integral. Consequently, this value must be zero, and the proof is complete.

The proof for *the case in which C is the boundary of a polygon* follows from the previous proof by subdividing the polygon into triangles (Fig. 564). The integral corresponding to each such triangle is zero. The sum of these integrals is equal to the integral over C, because we integrate along each segment of subdivision in both directions, the corresponding integrals cancel out in pairs, and we are left with the integral over C.

The case of a general simple closed path C can be reduced to the preceding one by inscribing in C a closed polygon P of chords, which approximates C "sufficiently accurately," and it can be shown that there is a polygon P such that the integral over P differs from that over C by less than any preassigned positive real number $\tilde{\epsilon}$, no matter how small. The details of this proof are somewhat involved and can be found in Ref. [D6] listed in App. 1. ■

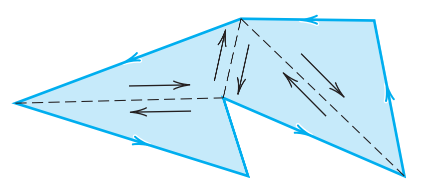


Fig. 564. Proof of Cauchy's integral theorem for a polygon

Section 15.1, page 674

PROOF OF THEOREM 4 Cauchy's Convergence Principle for Series

(a) In this proof we need two concepts and a theorem, which we list first.

1. A **bounded sequence** $s_1, s_2, \cdots$ is a sequence whose terms all lie in a disk of (sufficiently large, finite) radius K with center at the origin; thus $|s_n| < K$ for all n.

2. A **limit point** a of a sequence $s_1, s_2, \cdots$ is a point such that, given an $\epsilon > 0$, there are infinitely many terms satisfying $|s_n - a| < \epsilon$. (Note that this does *not* imply convergence, since there may still be infinitely many terms that do not lie within that circle of radius ϵ and center a.)

Example: $\frac{1}{4}, \frac{3}{4}, \frac{1}{8}, \frac{7}{8}, \frac{1}{16}, \frac{15}{16}, \cdots$ has the limit points 0 and 1 and diverges.

3. A bounded sequence in the complex plane has at least one limit point. (Bolzano–Weierstrass theorem; proof below. Recall that "sequence" always means *infinite* sequence.)

(b) We now turn to the actual proof that $z_1 + z_2 + \cdots$ converges if and only if, for every $\epsilon > 0$, we can find an N such that

$$|z_{n+1} + \cdots + z_{n+p}| < \epsilon \qquad \text{for every } n > N \text{ and } p = 1, 2, \cdots. \tag{1}$$

Here, by the definition of partial sums,

$$s_{n+p} - s_n = z_{n+1} + \cdots + z_{n+p}.$$

Writing $n + p = r$, we see from this that (1) is equivalent to

$$|s_r - s_n| < \epsilon \qquad \text{for all } r > N \text{ and } n > N. \tag{1*}$$

Suppose that $s_1, s_2, \cdots$ converges. Denote its limit by s. Then for a given $\epsilon > 0$ we can find an N such that

$$|s_n - s| < \frac{\epsilon}{2} \qquad \text{for every } n > N.$$

Hence, if $r > N$ and $n > N$, then by the triangle inequality (Sec. 13.2),

$$|s_r - s_n| = |(s_r - s) - (s_n - s)| \leqq |s_r - s| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon,$$

that is, (1*) holds.

(c) Conversely, assume that $s_1, s_2, \cdots$ satisfies (1*). We first prove that then the sequence must be bounded. Indeed, choose a fixed ϵ and a fixed $n = n_0 > N$ in (1*). Then (1*) implies that all s_r with $r > N$ lie in the disk of radius ϵ and center s_{n_0} and only *finitely many terms* $s_1, \cdots, s_N$ may not lie in this disk. Clearly, we can now find a circle so large that this disk and these finitely many terms all lie within this new circle. Hence the sequence is bounded. By the Bolzano–Weierstrass theorem, it has at least one limit point, call it s.

We now show that the sequence is convergent with the limit s. Let $\epsilon > 0$ be given. Then there is an N^* such that $|s_r - s_n| < \epsilon/2$ for all $r > N^*$ and $n > N^*$, by (1*). Also, by the definition of a limit point, $|s_n - s| < \epsilon/2$ for *infinitely many n*, so that we can find and fix an $n > N^*$ such that $|s_n - s| < \epsilon/2$. Together, for *every* $r > N^*$,

$$|s_r - s| = |(s_r - s_n) + (s_n - s)| \leqq |s_r - s_n| + |s_n - s| < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon;$$

that is, the sequence $s_1, s_2, \cdots$ is convergent with the limit s. ∎

THEOREM

Bolzano–Weierstrass Theorem[3]

A bounded infinite sequence $z_1, z_2, z_3, \cdots$ in the complex plane has at least one limit point.

PROOF It is obvious that we need both conditions: a finite sequence cannot have a limit point, and the sequence $1, 2, 3, \cdots$, which is infinite but not bounded, has no limit point. To prove the theorem, consider a bounded infinite sequence $z_1, z_2, \cdots$ and let K be such that $|z_n| < K$ for all n. If only finitely many values of the z_n are different, then, since the sequence is infinite, some number z must occur infinitely many times in the sequence, and, by definition, this number is a limit point of the sequence.

We may now turn to the case when the sequence contains infinitely many *different* terms. We draw a large square Q_0 that contains all z_n. We subdivide Q_0 into four congruent squares, which we number 1, 2, 3, 4. Clearly, at least one of these squares (each taken with its complete boundary) must contain infinitely many terms of the sequence. The square of this type with the lowest number (1, 2, 3, or 4) will be denoted by Q_1. This is

[3]BERNARD BOLZANO (1781–1848), Austrian mathematician and professor of religious studies, was a pioneer in the study of point sets, the foundation of analysis, and mathematical logic.

For Weierstrass, see Sec. 15.5.

the first step. In the next step we subdivide Q_1 into four congruent squares and select a square Q_2 by the same rule, and so on. This yields an infinite sequence of squares Q_0, $Q_1, Q_2, \cdots, Q_n, \cdots$ with the property that the side of Q_n approaches zero as n approaches infinity, and Q_m contains all Q_n with $n > m$. It is not difficult to see that the number which belongs to all these squares,[4] call it $z = a$, is a limit point of the sequence. In fact, given an $\epsilon > 0$, we can choose an N so large that the side of the square Q_N is less than ϵ and, since Q_N contains infinitely many z_n, we have $|z_n - a| < \epsilon$ for infinitely many n. This completes the proof. ■

Section 15.3, pages 688–689

PART (b) OF THE PROOF OF THEOREM 5

We have to show that

$$\sum_{n=2}^{\infty} a_n \left[\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1} \right]$$

$$= \sum_{n=2}^{\infty} a_n \, \Delta z \left[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n-1)z^{n-2} \right],$$

thus,

$$\frac{(z + \Delta z)^n - z^n}{\Delta z} - nz^{n-1}$$

$$= \Delta z \left[(z + \Delta z)^{n-2} + 2z(z + \Delta z)^{n-3} + \cdots + (n-1)z^{n-2} \right].$$

If we set $z + \Delta z = b$ and $z = a$, thus $\Delta z = b - a$, this becomes simply

$$\text{(7a)} \qquad \frac{b^n - a^n}{b - a} - na^{n-1} = (b - a)A_n \qquad (n = 2, 3, \cdots),$$

where A_n is the expression in the brackets on the right,

$$\text{(7b)} \qquad A_n = b^{n-2} + 2ab^{n-3} + 3a^2b^{n-4} + \cdots + (n-1)a^{n-2};$$

thus, $A_2 = 1$, $A_3 = b + 2a$, etc. We prove (7) by induction. When $n = 2$, then (7) holds, since then

$$\frac{b^2 - a^2}{b - a} - 2a = \frac{(b + a)(b - a)}{b - a} - 2a = b - a = (b - a)A_2.$$

Assuming that (7) holds for $n = k$, we show that it holds for $n = k + 1$. By adding and subtracting a term in the numerator and then dividing we first obtain

$$\frac{b^{k+1} - a^{k+1}}{b - a} = \frac{b^{k+1} - ba^k + ba^k - a^{k+1}}{b - a} = b\,\frac{b^k - a^k}{b - a} + a^k.$$

[4]The fact that such a unique number $z = a$ exists seems to be obvious, but it actually follows from an axiom of the real number system, the so-called *Cantor–Dedekind axiom:* see footnote 3 in App. A3.3.

By the induction hypothesis, the right side equals $b[(b-a)A_k + ka^{k-1}] + a^k$. Direct calculation shows that this is equal to

$$(b-a)\{bA_k + ka^{k-1}\} + aka^{k-1} + a^k.$$

From (7b) with $n = k$ we see that the expression in the braces $\{\cdots\}$ equals

$$b^{k-1} + 2ab^{k-2} + \cdots + (k-1)ba^{k-2} + ka^{k-1} = A_{k+1}.$$

Hence our result is

$$\frac{b^{k+1} - a^{k+1}}{b-a} = (b-a)A_{k+1} + (k+1)a^k.$$

Taking the last term to the left, we obtain (7) with $n = k + 1$. This proves (7) for any integer $n \geqq 2$ and completes the proof. ■

Section 18.2, page 763

ANOTHER PROOF OF THEOREM 1 *without the use of a harmonic conjugate*

We show that if $w = u + iv = f(z)$ is analytic and maps a domain D conformally onto a domain D^* and $\Phi^*(u, v)$ is harmonic in D^*, then

$$\Phi(x, y) = \Phi^*(u(x, y), v(x, y)) \tag{1}$$

is harmonic in D, that is, $\nabla^2\Phi = 0$ in D. We make no use of a harmonic conjugate of Φ^*, but use straightforward differentiation. By the chain rule,

$$\Phi_x = \Phi_u^* u_x + \Phi_v^* v_x.$$

We apply the chain rule again, underscoring the terms that will drop out when we form $\nabla^2\Phi$:

$$\Phi_{xx} = \underline{\Phi_u^* u_{xx}} + (\Phi_{uu}^* u_x + \underline{\Phi_{uv}^* v_x})u_x$$
$$+ \underline{\Phi_u^* v_{xx}} + (\underline{\Phi_{vu}^* u_x} + \Phi_{vv}^* v_x)v_x.$$

Φ_{yy} is the same with each x replaced by y. We form the sum $\nabla^2\Phi$. In it, $\Phi_{vu}^* = \Phi_{uv}^*$ is multiplied by

$$u_x v_x + u_y v_y$$

which is 0 by the Cauchy–Riemann equations. Also $\nabla^2 u = 0$ and $\nabla^2 v = 0$. There remains

$$\nabla^2\Phi = \Phi_{uu}^*(u_x^2 + u_y^2) + \Phi_{vv}^*(v_x^2 + v_y^2).$$

By the Cauchy–Riemann equations this becomes

$$\nabla^2\Phi = (\Phi_{uu}^* + \Phi_{vv}^*)(u_x^2 + v_x^2)$$

and is 0 since Φ^* is harmonic. ■

APPENDIX 5

Tables

For Tables of Laplace Transforms see Secs. 6.8 and 6.9.
For Tables of Fourier Transforms see Sec. 11.10.

If you have a Computer Algebra System (CAS), you may not need the present tables, but you may still find them convenient from time to time.

Table A1 Bessel Functions

For more extensive tables see Ref. [GenRef1] in App. 1.

x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$	x	$J_0(x)$	$J_1(x)$
0.0	1.0000	0.0000	3.0	−0.2601	0.3391	6.0	0.1506	−0.2767
0.1	0.9975	0.0499	3.1	−0.2921	0.3009	6.1	0.1773	−0.2559
0.2	0.9900	0.0995	3.2	−0.3202	0.2613	6.2	0.2017	−0.2329
0.3	0.9776	0.1483	3.3	−0.3443	0.2207	6.3	0.2238	−0.2081
0.4	0.9604	0.1960	3.4	−0.3643	0.1792	6.4	0.2433	−0.1816
0.5	0.9385	0.2423	3.5	−0.3801	0.1374	6.5	0.2601	−0.1538
0.6	0.9120	0.2867	3.6	−0.3918	0.0955	6.6	0.2740	−0.1250
0.7	0.8812	0.3290	3.7	−0.3992	0.0538	6.7	0.2851	−0.0953
0.8	0.8463	0.3688	3.8	−0.4026	0.0128	6.8	0.2931	−0.0652
0.9	0.8075	0.4059	3.9	−0.4018	−0.0272	6.9	0.2981	−0.0349
1.0	0.7652	0.4401	4.0	−0.3971	−0.0660	7.0	0.3001	−0.0047
1.1	0.7196	0.4709	4.1	−0.3887	−0.1033	7.1	0.2991	0.0252
1.2	0.6711	0.4983	4.2	−0.3766	−0.1386	7.2	0.2951	0.0543
1.3	0.6201	0.5220	4.3	−0.3610	−0.1719	7.3	0.2882	0.0826
1.4	0.5669	0.5419	4.4	−0.3423	−0.2028	7.4	0.2786	0.1096
1.5	0.5118	0.5579	4.5	−0.3205	−0.2311	7.5	0.2663	0.1352
1.6	0.4554	0.5699	4.6	−0.2961	−0.2566	7.6	0.2516	0.1592
1.7	0.3980	0.5778	4.7	−0.2693	−0.2791	7.7	0.2346	0.1813
1.8	0.3400	0.5815	4.8	−0.2404	−0.2985	7.8	0.2154	0.2014
1.9	0.2818	0.5812	4.9	−0.2097	−0.3147	7.9	0.1944	0.2192
2.0	0.2239	0.5767	5.0	−0.1776	−0.3276	8.0	0.1717	0.2346
2.1	0.1666	0.5683	5.1	−0.1443	−0.3371	8.1	0.1475	0.2476
2.2	0.1104	0.5560	5.2	−0.1103	−0.3432	8.2	0.1222	0.2580
2.3	0.0555	0.5399	5.3	−0.0758	−0.3460	8.3	0.0960	0.2657
2.4	0.0025	0.5202	5.4	−0.0412	−0.3453	8.4	0.0692	0.2708
2.5	−0.0484	0.4971	5.5	−0.0068	−0.3414	8.5	0.0419	0.2731
2.6	−0.0968	0.4708	5.6	0.0270	−0.3343	8.6	0.0146	0.2728
2.7	−0.1424	0.4416	5.7	0.0599	−0.3241	8.7	−0.0125	0.2697
2.8	−0.1850	0.4097	5.8	0.0917	−0.3110	8.8	−0.0392	0.2641
2.9	−0.2243	0.3754	5.9	0.1220	−0.2951	8.9	−0.0653	0.2559

$J_0(x) = 0$ for $x = 2.40483, 5.52008, 8.65373, 11.7915, 14.9309, 18.0711, 21.2116, 24.3525, 27.4935, 30.6346$

$J_1(x) = 0$ for $x = 3.83171, 7.01559, 10.1735, 13.3237, 16.4706, 19.6159, 22.7601, 25.9037, 29.0468, 32.1897$

Table A1 ***(continued)***

x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$	x	$Y_0(x)$	$Y_1(x)$
0.0	$(-\infty)$	$(-\infty)$	2.5	0.498	0.146	5.0	−0.309	0.148
0.5	−0.445	−1.471	3.0	0.377	0.325	5.5	−0.339	−0.024
1.0	0.088	−0.781	3.5	0.189	0.410	6.0	−0.288	−0.175
1.5	0.382	−0.412	4.0	−0.017	0.398	6.5	−0.173	−0.274
2.0	0.510	−0.107	4.5	−0.195	0.301	7.0	−0.026	−0.303

Table A2 **Gamma Function** [see (24) in App. A3.1]

α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$	α	$\Gamma(\alpha)$
1.00	1.000 000	1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384
1.02	0.988 844	1.22	0.913 106	1.42	0.886 356	1.62	0.895 924	1.82	0.936 845
1.04	0.978 438	1.24	0.908 521	1.44	0.885 805	1.64	0.898 642	1.84	0.942 612
1.06	0.968 744	1.26	0.904 397	1.46	0.885 604	1.66	0.901 668	1.86	0.948 687
1.08	0.959 725	1.28	0.900 718	1.48	0.885 747	1.68	0.905 001	1.88	0.955 071
1.10	0.951 351	1.30	0.897 471	1.50	0.886 227	1.70	0.908 639	1.90	0.961 766
1.12	0.943 590	1.32	0.894 640	1.52	0.887 039	1.72	0.912 581	1.92	0.968 774
1.14	0.936 416	1.34	0.892 216	1.54	0.888 178	1.74	0.916 826	1.94	0.976 099
1.16	0.929 803	1.36	0.890 185	1.56	0.889 639	1.76	0.921 375	1.96	0.983 743
1.18	0.923 728	1.38	0.888 537	1.58	0.891 420	1.78	0.926 227	1.98	0.991 708
1.20	0.918 169	1.40	0.887 264	1.60	0.893 515	1.80	0.931 384	2.00	1.000 000

Table A3 **Factorial Function and Its Logarithm with Base 10**

n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$	n	$n!$	$\log(n!)$
1	1	0.000 000	6	720	2.857 332	11	39 916 800	7.601 156
2	2	0.301 030	7	5 040	3.702 431	12	479 001 600	8.680 337
3	6	0.778 151	8	40 320	4.605 521	13	6 227 020 800	9.794 280
4	24	1.380 211	9	362 880	5.559 763	14	87 178 291 200	10.940 408
5	120	2.079 181	10	3 628 800	6.559 763	15	1 307 674 368 000	12.116 500

Table A4 **Error Function, Sine and Cosine Integrals** [see (35), (40), (42) in App. A3.1]

x	erf x	Si(x)	ci(x)	x	erf x	Si(x)	ci(x)
0.0	0.0000	0.0000	∞	2.0	0.9953	1.6054	−0.4230
0.2	0.2227	0.1996	1.0422	2.2	0.9981	1.6876	−0.3751
0.4	0.4284	0.3965	0.3788	2.4	0.9993	1.7525	−0.3173
0.6	0.6039	0.5881	0.0223	2.6	0.9998	1.8004	−0.2533
0.8	0.7421	0.7721	−0.1983	2.8	0.9999	1.8321	−0.1865
1.0	0.8427	0.9461	−0.3374	3.0	1.0000	1.8487	−0.1196
1.2	0.9103	1.1080	−0.4205	3.2	1.0000	1.8514	−0.0553
1.4	0.9523	1.2562	−0.4620	3.4	1.0000	1.8419	0.0045
1.6	0.9763	1.3892	−0.4717	3.6	1.0000	1.8219	0.0580
1.8	0.9891	1.5058	−0.4568	3.8	1.0000	1.7934	0.1038
2.0	0.9953	1.6054	−0.4230	4.0	1.0000	1.7582	0.1410

Table A5 Binomial Distribution

Probability function $f(x)$ [see (2), Sec. 24.7] and distribution function $F(x)$

n	x	$p = 0.1$		$p = 0.2$		$p = 0.3$		$p = 0.4$		$p = 0.5$	
		$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
		0.		**0.**		**0.**		**0.**		**0.**	
1	0	9000	0.9000	8000	0.8000	7000	0.7000	6000	0.6000	5000	0.5000
	1	1000	1.0000	2000	1.0000	3000	1.0000	4000	1.0000	5000	1.0000
2	0	8100	0.8100	6400	0.6400	4900	0.4900	3600	0.3600	2500	0.2500
	1	1800	0.9900	3200	0.9600	4200	0.9100	4800	0.8400	5000	0.7500
	2	0100	1.0000	0400	1.0000	0900	1.0000	1600	1.0000	2500	1.0000
3	0	7290	0.7290	5120	0.5120	3430	0.3430	2160	0.2160	1250	0.1250
	1	2430	0.9720	3840	0.8960	4410	0.7840	4320	0.6480	3750	0.5000
	2	0270	0.9990	0960	0.9920	1890	0.9730	2880	0.9360	3750	0.8750
	3	0010	1.0000	0080	1.0000	0270	1.0000	0640	1.0000	1250	1.0000
4	0	6561	0.6561	4096	0.4096	2401	0.2401	1296	0.1296	0625	0.0625
	1	2916	0.9477	4096	0.8192	4116	0.6517	3456	0.4752	2500	0.3125
	2	0486	0.9963	1536	0.9728	2646	0.9163	3456	0.8208	3750	0.6875
	3	0036	0.9999	0256	0.9984	0756	0.9919	1536	0.9744	2500	0.9375
	4	0001	1.0000	0016	1.0000	0081	1.0000	0256	1.0000	0625	1.0000
5	0	5905	0.5905	3277	0.3277	1681	0.1681	0778	0.0778	0313	0.0313
	1	3281	0.9185	4096	0.7373	3602	0.5282	2592	0.3370	1563	0.1875
	2	0729	0.9914	2048	0.9421	3087	0.8369	3456	0.6826	3125	0.5000
	3	0081	0.9995	0512	0.9933	1323	0.9692	2304	0.9130	3125	0.8125
	4	0005	1.0000	0064	0.9997	0284	0.9976	0768	0.9898	1563	0.9688
	5	0000	1.0000	0003	1.0000	0024	1.0000	0102	1.0000	0313	1.0000
6	0	5314	0.5314	2621	0.2621	1176	0.1176	0467	0.0467	0156	0.0156
	1	3543	0.8857	3932	0.6554	3025	0.4202	1866	0.2333	0938	0.1094
	2	0984	0.9841	2458	0.9011	3241	0.7443	3110	0.5443	2344	0.3438
	3	0146	0.9987	0819	0.9830	1852	0.9295	2765	0.8208	3125	0.6563
	4	0012	0.9999	0154	0.9984	0595	0.9891	1382	0.9590	2344	0.8906
	5	0001	1.0000	0015	0.9999	0102	0.9993	0369	0.9959	0938	0.9844
	6	0000	1.0000	0001	1.0000	0007	1.0000	0041	1.0000	0156	1.0000
7	0	4783	0.4783	2097	0.2097	0824	0.0824	0280	0.0280	0078	0.0078
	1	3720	0.8503	3670	0.5767	2471	0.3294	1306	0.1586	0547	0.0625
	2	1240	0.9743	2753	0.8520	3177	0.6471	2613	0.4199	1641	0.2266
	3	0230	0.9973	1147	0.9667	2269	0.8740	2903	0.7102	2734	0.5000
	4	0026	0.9998	0287	0.9953	0972	0.9712	1935	0.9037	2734	0.7734
	5	0002	1.0000	0043	0.9996	0250	0.9962	0774	0.9812	1641	0.9375
	6	0000	1.0000	0004	1.0000	0036	0.9998	0172	0.9984	0547	0.9922
	7	0000	1.0000	0000	1.0000	0002	1.0000	0016	1.0000	0078	1.0000
8	0	4305	0.4305	1678	0.1678	0576	0.0576	0168	0.0168	0039	0.0039
	1	3826	0.8131	3355	0.5033	1977	0.2553	0896	0.1064	0313	0.0352
	2	1488	0.9619	2936	0.7969	2965	0.5518	2090	0.3154	1094	0.1445
	3	0331	0.9950	1468	0.9437	2541	0.8059	2787	0.5941	2188	0.3633
	4	0046	0.9996	0459	0.9896	1361	0.9420	2322	0.8263	2734	0.6367
	5	0004	1.0000	0092	0.9988	0467	0.9887	1239	0.9502	2188	0.8555
	6	0000	1.0000	0011	0.9999	0100	0.9987	0413	0.9915	1094	0.9648
	7	0000	1.0000	0001	1.0000	0012	0.9999	0079	0.9993	0313	0.9961
	8	0000	1.0000	0000	1.0000	0001	1.0000	0007	1.0000	0039	1.0000

Table A6 Poisson Distribution

Probability function $f(x)$ [see (5), Sec. 24.7] and distribution function $F(x)$

x	$\mu = 0.1$		$\mu = 0.2$		$\mu = 0.3$		$\mu = 0.4$		$\mu = 0.5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	9048	0.9048	8187	0.8187	7408	0.7408	6703	0.6703	6065	0.6065
1	0905	0.9953	1637	0.9825	2222	0.9631	2681	0.9384	3033	0.9098
2	0045	0.9998	0164	0.9989	0333	0.9964	0536	0.9921	0758	0.9856
3	0002	1.0000	0011	0.9999	0033	0.9997	0072	0.9992	0126	0.9982
4	0000	1.0000	0001	1.0000	0003	1.0000	0007	0.9999	0016	0.9998
5							0001	1.0000	0002	1.0000

x	$\mu = 0.6$		$\mu = 0.7$		$\mu = 0.8$		$\mu = 0.9$		$\mu = 1$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	5488	0.5488	4966	0.4966	4493	0.4493	4066	0.4066	3679	0.3679
1	3293	0.8781	3476	0.8442	3595	0.8088	3659	0.7725	3679	0.7358
2	0988	0.9769	1217	0.9659	1438	0.9526	1647	0.9371	1839	0.9197
3	0198	0.9966	0284	0.9942	0383	0.9909	0494	0.9865	0613	0.9810
4	0030	0.9996	0050	0.9992	0077	0.9986	0111	0.9977	0153	0.9963
5	0004	1.0000	0007	0.9999	0012	0.9998	0020	0.9997	0031	0.9994
6			0001	1.0000	0002	1.0000	0003	1.0000	0005	0.9999
7									0001	1.0000

x	$\mu = 1.5$		$\mu = 2$		$\mu = 3$		$\mu = 4$		$\mu = 5$	
	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$	$f(x)$	$F(x)$
	0.		**0.**		**0.**		**0.**		**0.**	
0	2231	0.2231	1353	0.1353	0498	0.0498	0183	0.0183	0067	0.0067
1	3347	0.5578	2707	0.4060	1494	0.1991	0733	0.0916	0337	0.0404
2	2510	0.8088	2707	0.6767	2240	0.4232	1465	0.2381	0842	0.1247
3	1255	0.9344	1804	0.8571	2240	0.6472	1954	0.4335	1404	0.2650
4	0471	0.9814	0902	0.9473	1680	0.8153	1954	0.6288	1755	0.4405
5	0141	0.9955	0361	0.9834	1008	0.9161	1563	0.7851	1755	0.6160
6	0035	0.9991	0120	0.9955	0504	0.9665	1042	0.8893	1462	0.7622
7	0008	0.9998	0034	0.9989	0216	0.9881	0595	0.9489	1044	0.8666
8	0001	1.0000	0009	0.9998	0081	0.9962	0298	0.9786	0653	0.9319
9			0002	1.0000	0027	0.9989	0132	0.9919	0363	0.9682
10					0008	0.9997	0053	0.9972	0181	0.9863
11					0002	0.9999	0019	0.9991	0082	0.9945
12					0001	1.0000	0006	0.9997	0034	0.9980
13							0002	0.9999	0013	0.9993
14							0001	1.0000	0005	0.9998
15									0002	0.9999
16									0000	1.0000

Table A7 Normal Distribution

Values of the distribution function $\Phi(z)$ [see (3), Sec. 24.8]. $\Phi(-z) = 1 - \Phi(z)$

z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$	z	$\Phi(z)$
	0.		**0.**		**0.**		**0.**		**0.**		**0.**
0.01	5040	0.51	6950	1.01	8438	1.51	9345	2.01	9778	2.51	9940
0.02	5080	0.52	6985	1.02	8461	1.52	9357	2.02	9783	2.52	9941
0.03	5120	0.53	7019	1.03	8485	1.53	9370	2.03	9788	2.53	9943
0.04	5160	0.54	7054	1.04	8508	1.54	9382	2.04	9793	2.54	9945
0.05	5199	0.55	7088	1.05	8531	1.55	9394	2.05	9798	2.55	9946
0.06	5239	0.56	7123	1.06	8554	1.56	9406	2.06	9803	2.56	9948
0.07	5279	0.57	7157	1.07	8577	1.57	9418	2.07	9808	2.57	9949
0.08	5319	0.58	7190	1.08	8599	1.58	9429	2.08	9812	2.58	9951
0.09	5359	0.59	7224	1.09	8621	1.59	9441	2.09	9817	2.59	9952
0.10	5398	0.60	7257	1.10	8643	1.60	9452	2.10	9821	2.60	9953
0.11	5438	0.61	7291	1.11	8665	1.61	9463	2.11	9826	2.61	9955
0.12	5478	0.62	7324	1.12	8686	1.62	9474	2.12	9830	2.62	9956
0.13	5517	0.63	7357	1.13	8708	1.63	9484	2.13	9834	2.63	9957
0.14	5557	0.64	7389	1.14	8729	1.64	9495	2.14	9838	2.64	9959
0.15	5596	0.65	7422	1.15	8749	1.65	9505	2.15	9842	2.65	9960
0.16	5636	0.66	7454	1.16	8770	1.66	9515	2.16	9846	2.66	9961
0.17	5675	0.67	7486	1.17	8790	1.67	9525	2.17	9850	2.67	9962
0.18	5714	0.68	7517	1.18	8810	1.68	9535	2.18	9854	2.68	9963
0.19	5753	0.69	7549	1.19	8830	1.69	9545	2.19	9857	2.69	9964
0.20	5793	0.70	7580	1.20	8849	1.70	9554	2.20	9861	2.70	9965
0.21	5832	0.71	7611	1.21	8869	1.71	9564	2.21	9864	2.71	9966
0.22	5871	0.72	7642	1.22	8888	1.72	9573	2.22	9868	2.72	9967
0.23	5910	0.73	7673	1.23	8907	1.73	9582	2.23	9871	2.73	9968
0.24	5948	0.74	7704	1.24	8925	1.74	9591	2.24	9875	2.74	9969
0.25	5987	0.75	7734	1.25	8944	1.75	9599	2.25	9878	2.75	9970
0.26	6026	0.76	7764	1.26	8962	1.76	9608	2.26	9881	2.76	9971
0.27	6064	0.77	7794	1.27	8980	1.77	9616	2.27	9884	2.77	9972
0.28	6103	0.78	7823	1.28	8997	1.78	9625	2.28	9887	2.78	9973
0.29	6141	0.79	7852	1.29	9015	1.79	9633	2.29	9890	2.79	9974
0.30	6179	0.80	7881	1.30	9032	1.80	9641	2.30	9893	2.80	9974
0.31	6217	0.81	7910	1.31	9049	1.81	9649	2.31	9896	2.81	9975
0.32	6255	0.82	7939	1.32	9066	1.82	9656	2.32	9898	2.82	9976
0.33	6293	0.83	7967	1.33	9082	1.83	9664	2.33	9901	2.83	9977
0.34	6331	0.84	7995	1.34	9099	1.84	9671	2.34	9904	2.84	9977
0.35	6368	0.85	8023	1.35	9115	1.85	9678	2.35	9906	2.85	9978
0.36	6406	0.86	8051	1.36	9131	1.86	9686	2.36	9909	2.86	9979
0.37	6443	0.87	8078	1.37	9147	1.87	9693	2.37	9911	2.87	9979
0.38	6480	0.88	8106	1.38	9162	1.88	9699	2.38	9913	2.88	9980
0.39	6517	0.89	8133	1.39	9177	1.89	9706	2.39	9916	2.89	9981
0.40	6554	0.90	8159	1.40	9192	1.90	9713	2.40	9918	2.90	9981
0.41	6591	0.91	8186	1.41	9207	1.91	9719	2.41	9920	2.91	9982
0.42	6628	0.92	8212	1.42	9222	1.92	9726	2.42	9922	2.92	9982
0.43	6664	0.93	8238	1.43	9236	1.93	9732	2.43	9925	2.93	9983
0.44	6700	0.94	8264	1.44	9251	1.94	9738	2.44	9927	2.94	9984
0.45	6736	0.95	8289	1.45	9265	1.95	9744	2.45	9929	2.95	9984
0.46	6772	0.96	8315	1.46	9279	1.96	9750	2.46	9931	2.96	9985
0.47	6808	0.97	8340	1.47	9292	1.97	9756	2.47	9932	2.97	9985
0.48	6844	0.98	8365	1.48	9306	1.98	9761	2.48	9934	2.98	9986
0.49	6879	0.99	8389	1.49	9319	1.99	9767	2.49	9936	2.99	9986
0.50	6915	1.00	8413	1.50	9332	2.00	9772	2.50	9938	3.00	9987

Table A8 Normal Distribution

Values of z for given values of $\Phi(z)$ [see (3), Sec. 24.8] and $D(z) = \Phi(z) - \Phi(-z)$
Example: $z = 0.279$ if $\Phi(z) = 61\%$; $z = 0.860$ if $D(z) = 61\%$.

%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$	%	$z(\Phi)$	$z(D)$
1	−2.326	0.013	41	−0.228	0.539	81	0.878	1.311
2	−2.054	0.025	42	−0.202	0.553	82	0.915	1.341
3	−1.881	0.038	43	−0.176	0.568	83	0.954	1.372
4	−1.751	0.050	44	−0.151	0.583	84	0.994	1.405
5	−1.645	0.063	45	−0.126	0.598	85	1.036	1.440
6	−1.555	0.075	46	−0.100	0.613	86	1.080	1.476
7	−1.476	0.088	47	−0.075	0.628	87	1.126	1.514
8	−1.405	0.100	48	−0.050	0.643	88	1.175	1.555
9	−1.341	0.113	49	−0.025	0.659	89	1.227	1.598
10	−1.282	0.126	50	0.000	0.674	90	1.282	1.645
11	−1.227	0.138	51	0.025	0.690	91	1.341	1.695
12	−1.175	0.151	52	0.050	0.706	92	1.405	1.751
13	−1.126	0.164	53	0.075	0.722	93	1.476	1.812
14	−1.080	0.176	54	0.100	0.739	94	1.555	1.881
15	−1.036	0.189	55	0.126	0.755	95	1.645	1.960
16	−0.994	0.202	56	0.151	0.772	96	1.751	2.054
17	−0.954	0.215	57	0.176	0.789	97	1.881	2.170
18	−0.915	0.228	58	0.202	0.806	97.5	1.960	2.241
19	−0.878	0.240	59	0.228	0.824	98	2.054	2.326
20	−0.842	0.253	60	0.253	0.842	99	2.326	2.576
21	−0.806	0.266	61	0.279	0.860	99.1	2.366	2.612
22	−0.772	0.279	62	0.305	0.878	99.2	2.409	2.652
23	−0.739	0.292	63	0.332	0.896	99.3	2.457	2.697
24	−0.706	0.305	64	0.358	0.915	99.4	2.512	2.748
25	−0.674	0.319	65	0.385	0.935	99.5	2.576	2.807
26	−0.643	0.332	66	0.412	0.954	99.6	2.652	2.878
27	−0.613	0.345	67	0.440	0.974	99.7	2.748	2.968
28	−0.583	0.358	68	0.468	0.994	99.8	2.878	3.090
29	−0.553	0.372	69	0.496	1.015	99.9	3.090	3.291
30	−0.524	0.385	70	0.524	1.036			
31	−0.496	0.399	71	0.553	1.058	99.91	3.121	3.320
32	−0.468	0.412	72	0.583	1.080	99.92	3.156	3.353
33	−0.440	0.426	73	0.613	1.103	99.93	3.195	3.390
34	−0.412	0.440	74	0.643	1.126	99.94	3.239	3.432
35	−0.385	0.454	75	0.674	1.150	99.95	3.291	3.481
36	−0.358	0.468	76	0.706	1.175	99.96	3.353	3.540
37	−0.332	0.482	77	0.739	1.200	99.97	3.432	3.615
38	−0.305	0.496	78	0.772	1.227	99.98	3.540	3.719
39	−0.279	0.510	79	0.806	1.254	99.99	3.719	3.891
40	−0.253	0.524	80	0.842	1.282			

Table A9 ***t*-Distribution**

Values of z for given values of the distribution function $F(z)$ (see (8) in Sec. 25.3). Example: For 9 degrees of freedom, $z = 1.83$ when $F(z) = 0.95$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.32	0.29	0.28	0.27	0.27	0.26	0.26	0.26	0.26	0.26
0.7	0.73	0.62	0.58	0.57	0.56	0.55	0.55	0.55	0.54	0.54
0.8	1.38	1.06	0.98	0.94	0.92	0.91	0.90	0.89	0.88	0.88
0.9	3.08	1.89	1.64	1.53	1.48	1.44	1.41	1.40	1.38	1.37
0.95	6.31	2.92	2.35	2.13	2.02	1.94	1.89	1.86	1.83	1.81
0.975	12.7	4.30	3.18	2.78	2.57	2.45	2.36	2.31	2.26	2.23
0.99	31.8	6.96	4.54	3.75	3.36	3.14	3.00	2.90	2.82	2.76
0.995	63.7	9.92	5.84	4.60	4.03	3.71	3.50	3.36	3.25	3.17
0.999	318.3	22.3	10.2	7.17	5.89	5.21	4.79	4.50	4.30	4.14

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26	0.26
0.7	0.54	0.54	0.54	0.54	0.54	0.54	0.53	0.53	0.53	0.53
0.8	0.88	0.87	0.87	0.87	0.87	0.86	0.86	0.86	0.86	0.86
0.9	1.36	1.36	1.35	1.35	1.34	1.34	1.33	1.33	1.33	1.33
0.95	1.80	1.78	1.77	1.76	1.75	1.75	1.74	1.73	1.73	1.72
0.975	2.20	2.18	2.16	2.14	2.13	2.12	2.11	2.10	2.09	2.09
0.99	2.72	2.68	2.65	2.62	2.60	2.58	2.57	2.55	2.54	2.53
0.995	3.11	3.05	3.01	2.98	2.95	2.92	2.90	2.88	2.86	2.85
0.999	4.02	3.93	3.85	3.79	3.73	3.69	3.65	3.61	3.58	3.55

$F(z)$	Number of Degrees of Freedom									
	22	24	26	28	30	40	50	100	200	∞
0.5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.6	0.26	0.26	0.26	0.26	0.26	0.26	0.25	0.25	0.25	0.25
0.7	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.52
0.8	0.86	0.86	0.86	0.85	0.85	0.85	0.85	0.85	0.84	0.84
0.9	1.32	1.32	1.31	1.31	1.31	1.30	1.30	1.29	1.29	1.28
0.95	1.72	1.71	1.71	1.70	1.70	1.68	1.68	1.66	1.65	1.65
0.975	2.07	2.06	2.06	2.05	2.04	2.02	2.01	1.98	1.97	1.96
0.99	2.51	2.49	2.48	2.47	2.46	2.42	2.40	2.36	2.35	2.33
0.995	2.82	2.80	2.78	2.76	2.75	2.70	2.68	2.63	2.60	2.58
0.999	3.50	3.47	3.43	3.41	3.39	3.31	3.26	3.17	3.13	3.09

Table A10 Chi-square Distribution

Values of x for given values of the distribution function $F(z)$ (see Sec. 25.3 before (17)). Example: For 3 degrees of freedom, $z = 11.34$ when $F(z) = 0.99$.

$F(z)$	Number of Degrees of Freedom									
	1	2	3	4	5	6	7	8	9	10
0.005	0.00	0.01	0.07	0.21	0.41	0.68	0.99	1.34	1.73	2.16
0.01	0.00	0.02	0.11	0.30	0.55	0.87	1.24	1.65	2.09	2.56
0.025	0.00	0.05	0.22	0.48	0.83	1.24	1.69	2.18	2.70	3.25
0.05	0.00	0.10	0.35	0.71	1.15	1.64	2.17	2.73	3.33	3.94
0.95	3.84	5.99	7.81	9.49	11.07	12.59	14.07	15.51	16.92	18.31
0.975	5.02	7.38	9.35	11.14	12.83	14.45	16.01	17.53	19.02	20.48
0.99	6.63	9.21	11.34	13.28	15.09	16.81	18.48	20.09	21.67	23.21
0.995	7.88	10.60	12.84	14.86	16.75	18.55	20.28	21.95	23.59	25.19

$F(z)$	Number of Degrees of Freedom									
	11	12	13	14	15	16	17	18	19	20
0.005	2.60	3.07	3.57	4.07	4.60	5.14	5.70	6.26	6.84	7.43
0.01	3.05	3.57	4.11	4.66	5.23	5.81	6.41	7.01	7.63	8.26
0.025	3.82	4.40	5.01	5.63	6.26	6.91	7.56	8.23	8.91	9.59
0.05	4.57	5.23	5.89	6.57	7.26	7.96	8.67	9.39	10.12	10.85
0.95	19.68	21.03	22.36	23.68	25.00	26.30	27.59	28.87	30.14	31.41
0.975	21.92	23.34	24.74	26.12	27.49	28.85	30.19	31.53	32.85	34.17
0.99	24.72	26.22	27.69	29.14	30.58	32.00	33.41	34.81	36.19	37.57
0.995	26.76	28.30	29.82	31.32	32.80	34.27	35.72	37.16	38.58	40.00

$F(z)$	Number of Degrees of Freedom									
	21	22	23	24	25	26	27	28	29	30
0.005	8.0	8.6	9.3	9.9	10.5	11.2	11.8	12.5	13.1	13.8
0.01	8.9	9.5	10.2	10.9	11.5	12.2	12.9	13.6	14.3	15.0
0.025	10.3	11.0	11.7	12.4	13.1	13.8	14.6	15.3	16.0	16.8
0.05	11.6	12.3	13.1	13.8	14.6	15.4	16.2	16.9	17.7	18.5
0.95	32.7	33.9	35.2	36.4	37.7	38.9	40.1	41.3	42.6	43.8
0.975	35.5	36.8	38.1	39.4	40.6	41.9	43.2	44.5	45.7	47.0
0.99	38.9	40.3	41.6	43.0	44.3	45.6	47.0	48.3	49.6	50.9
0.995	41.4	42.8	44.2	45.6	46.9	48.3	49.6	51.0	52.3	53.7

$F(z)$	Number of Degrees of Freedom							
	40	50	60	70	80	90	100	> 100 (Approximation)
0.005	20.7	28.0	35.5	43.3	51.2	59.2	67.3	$\frac{1}{2}(h - 2.58)^2$
0.01	22.2	29.7	37.5	45.4	53.5	61.8	70.1	$\frac{1}{2}(h - 2.33)^2$
0.025	24.4	32.4	40.5	48.8	57.2	65.6	74.2	$\frac{1}{2}(h - 1.96)^2$
0.05	26.5	34.8	43.2	51.7	60.4	69.1	77.9	$\frac{1}{2}(h - 1.64)^2$
0.95	55.8	67.5	79.1	90.5	101.9	113.1	124.3	$\frac{1}{2}(h + 1.64)^2$
0.975	59.3	71.4	83.3	95.0	106.6	118.1	129.6	$\frac{1}{2}(h + 1.96)^2$
0.99	63.7	76.2	88.4	100.4	112.3	124.1	135.8	$\frac{1}{2}(h + 2.33)^2$
0.995	66.8	79.5	92.0	104.2	116.3	128.3	140.2	$\frac{1}{2}(h + 2.58)^2$

In the last column, $h = \sqrt{2m - 1}$, where m is the number of degrees of freedom.

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom**

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

Example: For (7, 4) d.f., $z = 6.09$ if $F(z) = 0.95$.

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	161	200	216	225	230	234	237	239	241
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21
32	4.15	3.29	2.90	2.67	2.51	2.40	2.31	2.24	2.19
34	4.13	3.28	2.88	2.65	2.49	2.38	2.29	2.23	2.17
36	4.11	3.26	2.87	2.63	2.48	2.36	2.28	2.21	2.15
38	4.10	3.24	2.85	2.62	2.46	2.35	2.26	2.19	2.14
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12
50	4.03	3.18	2.79	2.56	2.40	2.29	2.20	2.13	2.07
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04
70	3.98	3.13	2.74	2.50	2.35	2.23	2.14	2.07	2.02
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	2.00
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.97
150	3.90	3.06	2.66	2.43	2.27	2.16	2.07	2.00	1.94
200	3.89	3.04	2.65	2.42	2.26	2.14	2.06	1.98	1.93
1000	3.85	3.00	2.61	2.38	2.22	2.11	2.02	1.95	1.89
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.95**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	242	246	248	250	251	252	253	254
2	19.4	19.4	19.4	19.5	19.5	19.5	19.5	19.5
3	8.79	8.70	8.66	8.62	8.59	8.58	8.55	8.53
4	5.96	5.86	5.80	5.75	5.72	5.70	5.66	5.63
5	4.74	4.62	4.56	4.50	4.46	4.44	4.41	4.37
6	4.06	3.94	3.87	3.81	3.77	3.75	3.71	3.67
7	3.64	3.51	3.44	3.38	3.34	3.32	3.27	3.23
8	3.35	3.22	3.15	3.08	3.04	3.02	2.97	2.93
9	3.14	3.01	2.94	2.86	2.83	2.80	2.76	2.71
10	2.98	2.85	2.77	2.70	2.66	2.64	2.59	2.54
11	2.85	2.72	2.65	2.57	2.53	2.51	2.46	2.40
12	2.75	2.62	2.54	2.47	2.43	2.40	2.35	2.30
13	2.67	2.53	2.46	2.38	2.34	2.31	2.26	2.21
14	2.60	2.46	2.39	2.31	2.27	2.24	2.19	2.13
15	2.54	2.40	2.33	2.25	2.20	2.18	2.12	2.07
16	2.49	2.35	2.28	2.19	2.15	2.12	2.07	2.01
17	2.45	2.31	2.23	2.15	2.10	2.08	2.02	1.96
18	2.41	2.27	2.19	2.11	2.06	2.04	1.98	1.92
19	2.38	2.23	2.16	2.07	2.03	2.00	1.94	1.88
20	2.35	2.20	2.12	2.04	1.99	1.97	1.91	1.84
22	2.30	2.15	2.07	1.98	1.94	1.91	1.85	1.78
24	2.25	2.11	2.03	1.94	1.89	1.86	1.80	1.73
26	2.22	2.07	1.99	1.90	1.85	1.82	1.76	1.69
28	2.19	2.04	1.96	1.87	1.82	1.79	1.73	1.65
30	2.16	2.01	1.93	1.84	1.79	1.76	1.70	1.62
32	2.14	1.99	1.91	1.82	1.77	1.74	1.67	1.59
34	2.12	1.97	1.89	1.80	1.75	1.71	1.65	1.57
36	2.11	1.95	1.87	1.78	1.73	1.69	1.62	1.55
38	2.09	1.94	1.85	1.76	1.71	1.68	1.61	1.53
40	2.08	1.92	1.84	1.74	1.69	1.66	1.59	1.51
50	2.03	1.87	1.78	1.69	1.63	1.60	1.52	1.44
60	1.99	1.84	1.75	1.65	1.59	1.56	1.48	1.39
70	1.97	1.81	1.72	1.62	1.57	1.53	1.45	1.35
80	1.95	1.79	1.70	1.60	1.54	1.51	1.43	1.32
90	1.94	1.78	1.69	1.59	1.53	1.49	1.41	1.30
100	1.93	1.77	1.68	1.57	1.52	1.48	1.39	1.28
150	1.89	1.73	1.64	1.54	1.48	1.44	1.34	1.22
200	1.88	1.72	1.62	1.52	1.46	1.41	1.32	1.19
1000	1.84	1.68	1.58	1.47	1.41	1.36	1.26	1.08
∞	1.83	1.67	1.57	1.46	1.39	1.35	1.24	1.00

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 5$	$m = 6$	$m = 7$	$m = 8$	$m = 9$
1	4052	4999	5403	5625	5764	5859	5928	5981	6022
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07
32	7.50	5.34	4.46	3.97	3.65	3.43	3.26	3.13	3.02
34	7.44	5.29	4.42	3.93	3.61	3.39	3.22	3.09	2.98
36	7.40	5.25	4.38	3.89	3.57	3.35	3.18	3.05	2.95
38	7.35	5.21	4.34	3.86	3.54	3.32	3.15	3.02	2.92
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89
50	7.17	5.06	4.20	3.72	3.41	3.19	3.02	2.89	2.78
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72
70	7.01	4.92	4.07	3.60	3.29	3.07	2.91	2.78	2.67
80	6.96	4.88	4.04	3.56	3.26	3.04	2.87	2.74	2.64
90	6.93	4.85	4.01	3.54	3.23	3.01	2.84	2.72	2.61
100	6.90	4.82	3.98	3.51	3.21	2.99	2.82	2.69	2.59
150	6.81	4.75	3.91	3.45	3.14	2.92	2.76	2.63	2.53
200	6.76	4.71	3.88	3.41	3.11	2.89	2.73	2.60	2.50
1000	6.66	4.63	3.80	3.34	3.04	2.82	2.66	2.53	2.43
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41

Table A11 ***F*-Distribution with (*m*, *n*) Degrees of Freedom *(continued)***

Values of z for which the distribution function $F(z)$ [see (13), Sec. 25.4] has the value **0.99**

n	$m = 10$	$m = 15$	$m = 20$	$m = 30$	$m = 40$	$m = 50$	$m = 100$	∞
1	6056	6157	6209	6261	6287	6303	6334	6366
2	99.4	99.4	99.4	99.5	99.5	99.5	99.5	99.5
3	27.2	26.9	26.7	26.5	26.4	26.4	26.2	26.1
4	14.5	14.2	14.0	13.8	13.7	13.7	13.6	13.5
5	10.1	9.72	9.55	9.38	9.29	9.24	9.13	9.02
6	7.87	7.56	7.40	7.23	7.14	7.09	6.99	6.88
7	6.62	6.31	6.16	5.99	5.91	5.86	5.75	5.65
8	5.81	5.52	5.36	5.20	5.12	5.07	4.96	4.86
9	5.26	4.96	4.81	4.65	4.57	4.52	4.42	4.31
10	4.85	4.56	4.41	4.25	4.17	4.12	4.01	3.91
11	4.54	4.25	4.10	3.94	3.86	3.81	3.71	3.60
12	4.30	4.01	3.86	3.70	3.62	3.57	3.47	3.36
13	4.10	3.82	3.66	3.51	3.43	3.38	3.27	3.17
14	3.94	3.66	3.51	3.35	3.27	3.22	3.11	3.00
15	3.80	3.52	3.37	3.21	3.13	3.08	2.98	2.87
16	3.69	3.41	3.26	3.10	3.02	2.97	2.86	2.75
17	3.59	3.31	3.16	3.00	2.92	2.87	2.76	2.65
18	3.51	3.23	3.08	2.92	2.84	2.78	2.68	2.57
19	3.43	3.15	3.00	2.84	2.76	2.71	2.60	2.49
20	3.37	3.09	2.94	2.78	2.69	2.64	2.54	2.42
22	3.26	2.98	2.83	2.67	2.58	2.53	2.42	2.31
24	3.17	2.89	2.74	2.58	2.49	2.44	2.33	2.21
26	3.09	2.81	2.66	2.50	2.42	2.36	2.25	2.13
28	3.03	2.75	2.60	2.44	2.35	2.30	2.19	2.06
30	2.98	2.70	2.55	2.39	2.30	2.25	2.13	2.01
32	2.93	2.65	2.50	2.34	2.25	2.20	2.08	1.96
34	2.89	2.61	2.46	2.30	2.21	2.16	2.04	1.91
36	2.86	2.58	2.43	2.26	2.18	2.12	2.00	1.87
38	2.83	2.55	2.40	2.23	2.14	2.09	1.97	1.84
40	2.80	2.52	2.37	2.20	2.11	2.06	1.94	1.80
50	2.70	2.42	2.27	2.10	2.01	1.95	1.82	1.68
60	2.63	2.35	2.20	2.03	1.94	1.88	1.75	1.60
70	2.59	2.31	2.15	1.98	1.89	1.83	1.70	1.54
80	2.55	2.27	2.12	1.94	1.85	1.79	1.65	1.49
90	2.52	2.24	2.09	1.92	1.82	1.76	1.62	1.46
100	2.50	2.22	2.07	1.89	1.80	1.74	1.60	1.43
150	2.44	2.16	2.00	1.83	1.73	1.66	1.52	1.33
200	2.41	2.13	1.97	1.79	1.69	1.63	1.48	1.28
1000	2.34	2.06	1.90	1.72	1.61	1.54	1.38	1.11
∞	2.32	2.04	1.88	1.70	1.59	1.52	1.36	1.00

Table A12 Distribution Function $F(x) = P(T \leqq x)$ of the Random Variable T in Section 25.8

x	$n = 3$
	0.
0	167
1	500

x	$n = 4$
	0.
0	042
1	167
2	375

x	$n = 5$
	0.
0	008
1	042
2	117
3	242
4	408

x	$n = 6$
	0.
0	001
1	008
2	028
3	068
4	136
5	235
6	360
7	500

x	$n = 7$
	0.
1	001
2	005
3	015
4	035
5	068
6	119
7	191
8	281
9	386
10	500

x	$n = 8$
	0.
2	001
3	003
4	007
5	016
6	031
7	054
8	089
9	138
10	199
11	274
12	360
13	452

x	$n = 9$
	0.
4	001
5	003
6	006
7	012
8	022
9	038
10	060
11	090
12	130
13	179
14	238
15	306
16	381
17	460

x	$n = 10$
	0.
6	001
7	002
8	005
9	008
10	014
11	023
12	036
13	054
14	078
15	108
16	146
17	190
18	242
19	300
20	364
21	431
22	500

x	$n = 11$
	0.
8	001
9	002
10	003
11	005
12	008
13	013
14	020
15	030
16	043
17	060
18	082
19	109
20	141
21	179
22	223
23	271
24	324
25	381
26	440
27	500

x	$n = 20$
	0.
50	001
51	002
52	002
53	003
54	004
55	005
56	006
57	007
58	008
59	010
60	012
61	014
62	017
63	020
64	023
65	027
66	032
67	037
68	043
69	049
70	056
71	064
72	073
73	082
74	093
75	104
76	117
77	130
78	144
79	159
80	176
81	193
82	211
83	230
84	250
85	271
86	293
87	315
88	339
89	362
90	387
91	411
92	436
93	462
94	487

x	$n = 19$
	0.
43	001
44	002
45	002
46	003
47	003
48	004
49	005
50	006
51	008
52	010
53	012
54	014
55	017
56	021
57	025
58	029
59	034
60	040
61	047
62	054
63	062
64	072
65	082
66	093
67	105
68	119
69	133
70	149
71	166
72	184
73	203
74	223
75	245
76	267
77	290
78	314
79	339
80	365
81	391
82	418
83	445
84	473
85	500

x	$n = 18$
	0.
38	001
39	002
40	003
41	003
42	004
43	005
44	007
45	009
46	011
47	013
48	016
49	020
50	024
51	029
52	034
53	041
54	048
55	056
56	066
57	076
58	088
59	100
60	115
61	130
62	147
63	165
64	184
65	205
66	227
67	250
68	275
69	300
70	327
71	354
72	383
73	411
74	441
75	470
76	500

x	$n = 17$
	0.
32	001
33	002
34	002
35	003
36	004
37	005
38	007
39	009
40	011
41	014
42	017
43	021
44	026
45	032
46	038
47	046
48	054
49	064
50	076
51	088
52	102
53	118
54	135
55	154
56	174
57	196
58	220
59	245
60	271
61	299
62	328
63	358
64	388
65	420
66	452
67	484

x	$n = 16$
	0.
27	001
28	002
29	002
30	003
31	004
32	006
33	008
34	010
35	013
36	016
37	021
38	026
39	032
40	039
41	048
42	058
43	070
44	083
45	097
46	114
47	133
48	153
49	175
50	199
51	225
52	253
53	282
54	313
55	345
56	378
57	412
58	447
59	482

x	$n = 15$
	0.
23	001
24	002
25	003
26	004
27	006
28	008
29	010
30	014
31	018
32	023
33	029
34	037
35	046
36	057
37	070
38	084
39	101
40	120
41	141
42	164
43	190
44	218
45	248
46	279
47	313
48	349
49	385
50	423
51	461
52	500

x	$n = 14$
	0.
18	001
19	002
20	002
21	003
22	005
23	007
24	010
25	013
26	018
27	024
28	031
29	040
30	051
31	063
32	079
33	096
34	117
35	140
36	165
37	194
38	225
39	259
40	295
41	334
42	374
43	415
44	457
45	500

x	$n = 13$
	0.
14	001
15	001
16	002
17	003
18	005
19	007
20	011
21	015
22	021
23	029
24	038
25	050
26	064
27	082
28	102
29	126
30	153
31	184
32	218
33	255
34	295
35	338
36	383
37	429
38	476

x	$n = 12$
	0.
11	001
12	002
13	003
14	004
15	007
16	010
17	016
18	022
19	031
20	043
21	058
22	076
23	098
24	125
25	155
26	190
27	230
28	273
29	319
30	369
31	420
32	473

INDEX

PHOTO CREDITS

Part A Opener: © Denis Jr. Tangney/iStockphoto
Part B Opener: © Jill Fromer/iStockphoto
Part C Opener: © Science Photo Library/Photo Researchers, Inc
Part D Opener: © Rafa Irusta/iStockphoto
Part E Opener: © Alberto Pomares/iStockphoto
Chapter 19, Figure 437: © Eddie Gerald/Alamy
Part F Opener: © Rainer Plendl/iStockphoto
Part G Opener: © Sean Locke/iStockphoto
Appendix 1 Opener: © Ricardo De Mattos/iStockphoto
Appendix 2 Opener: © joel-t/iStockphoto
Appendix 3 Opener: © Luke Daniek/iStockphoto
Appendix 4 Opener: © Andrey Prokhorov/iStockphoto
Appendix 5 Opener: © Pedro Castellano/iStockphoto

Some Constants

$$e = 2.71828\ 18284\ 59045\ 23536$$
$$\sqrt{e} = 1.64872\ 12707\ 00128\ 14685$$
$$e^2 = 7.38905\ 60989\ 30650\ 22723$$

$$\pi = 3.14159\ 26535\ 89793\ 23846$$
$$\pi^2 = 9.86960\ 44010\ 89358\ 61883$$
$$\sqrt{\pi} = 1.77245\ 38509\ 05516\ 02730$$

$$\log_{10} \pi = 0.49714\ 98726\ 94133\ 85435$$
$$\ln \pi = 1.14472\ 98858\ 49400\ 17414$$
$$\log_{10} e = 0.43429\ 44819\ 03251\ 82765$$
$$\ln 10 = 2.30258\ 50929\ 94045\ 68402$$

$$\sqrt{2} = 1.41421\ 35623\ 73095\ 04880$$
$$\sqrt[3]{2} = 1.25992\ 10498\ 94873\ 16477$$
$$\sqrt{3} = 1.73205\ 08075\ 68877\ 29353$$
$$\sqrt[3]{3} = 1.44224\ 95703\ 07408\ 38232$$
$$\ln 2 = 0.69314\ 71805\ 59945\ 30942$$
$$\ln 3 = 1.09861\ 22886\ 68109\ 69140$$

$$\gamma = 0.57721\ 56649\ 01532\ 86061$$
$$\ln \gamma = -0.54953\ 93129\ 81644\ 82234$$
(see Sec. 5.6)
$$1^\circ = 0.01745\ 32925\ 19943\ 29577 \text{ rad}$$
$$1 \text{ rad} = 57.29577\ 95130\ 82320\ 87680^\circ$$
$$= 57^\circ 17' 44.806''$$

Polar Coordinates

$$x = r\cos\theta \qquad y = r\sin\theta$$
$$r = \sqrt{x^2 + y^2} \qquad \tan\theta = \frac{y}{x}$$
$$dx\,dy = r\,dr\,d\theta$$

Series

$$\frac{1}{1-x} = \sum_{m=0}^{\infty} x^m \qquad (|x| < 1)$$
$$e^x = \sum_{m=0}^{\infty} \frac{x^m}{m!}$$
$$\sin x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{(2m+1)!}$$
$$\cos x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m}}{(2m)!}$$
$$\ln(1-x) = -\sum_{m=1}^{\infty} \frac{x^m}{m} \qquad (|x| < 1)$$
$$\arctan x = \sum_{m=0}^{\infty} \frac{(-1)^m x^{2m+1}}{2m+1} \qquad (|x| < 1)$$

Greek Alphabet

α	Alpha	ν	Nu
β	Beta	ξ	Xi
γ, Γ	Gamma	o	Omicron
δ, Δ	Delta	π	Pi
ϵ, ε	Epsilon	ρ	Rho
ζ	Zeta	σ, Σ	Sigma
η	Eta	τ	Tau
θ, ϑ, Θ	Theta	υ, Υ	Upsilon
ι	Iota	ϕ, φ, Φ	Phi
κ	Kappa	χ	Chi
λ, Λ	Lambda	ψ, Ψ	Psi
μ	Mu	ω, Ω	Omega

Vectors

$$\mathbf{a}\cdot\mathbf{b} = a_1 b_1 + a_2 b_2 + a_3 b_3$$
$$\mathbf{a}\times\mathbf{b} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}$$
$$\text{grad } f = \nabla f = \frac{\partial f}{\partial x}\mathbf{i} + \frac{\partial f}{\partial y}\mathbf{j} + \frac{\partial f}{\partial z}\mathbf{k}$$
$$\text{div } \mathbf{v} = \nabla\cdot\mathbf{v} = \frac{\partial v_1}{\partial x} + \frac{\partial v_2}{\partial y} + \frac{\partial v_3}{\partial z}$$
$$\text{curl } \mathbf{v} = \nabla\times\mathbf{v} = \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ v_1 & v_2 & v_3 \end{vmatrix}$$